Spring 2011

Understanding Data Within The Enterprise To Control Or Reduce Data Storage Growth

Anthony Daubenmerkl
Regis University

Regis University
College for Professional Studies Graduate Programs
Final Project/Thesis

Disclaimer

UNDERSTANDING DATA WITHIN THE ENTERPRISE TO CONTROL OR REDUCE DATA STORAGE GROWTH

A THESIS

SUBMITTED ON THE FIRST OF MARCH, 2011

TO THE DEPARTMENT OF INFORMATION SYSTEMS

OF THE SCHOOL OF COMPUTER & INFORMATION SCIENCES

OF REGIS UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF MASTER OF SCIENCE IN INFORMATION TECHNOLOGY MANAGEMENT

BY

Anthony Daubenmerkl

APPROVALS

Robert Bowles, Thesis Advisor

Shari Plantz-Masters

Robert T. Mason

Abstract

Data is collected from a number of entities within the business environment; it is the essential component that guides the decision making process. Data management is a topic that continues to be discussed as our ability to collect and store data expands. The purpose of this study was to determine whether the data, in the context of the enterprise and its management, was understood allowing best practice data collection and storage policies to be defined. Data collection and storage are tightly coupled in the discussion of storage growth. The literature review showed that a number of data management lifecycles and solutions exist that support storage management policies, but do not focus upon controlling storage growth. Policies that affect or control storage growth do not necessarily reside within the application. This study deployed a survey that collected data which defined the problem area within the business enterprise. The survey data revealed that management does not understand the cost of storage nor data value. Furthermore, it was determined that the preferred method within the enterprise to handle the growing data was to expand storage. In addition to the survey, two interviews were used to substantiate the findings making the study a mixed method approach. The scope of the study was to define or provide a method to define best practice policies that would reduce or control storage growth. The findings have led to the development of the storage quadrant framework that provides the enterprise a method to define best practice policies.

Table of Contents

List of Figures

List of Tables

Chapter 1-Introduction

As technology has progressed, so has our ability to find new methods of data collection. Data exists everywhere. Within the enterprise, data has become an asset and for some, such as Nexus Lexus a lucrative business. Dependency upon the data stored perform valuable analytics has become an essential part of business activity. The value of data from all areas of the business has driven the expansion and growth of storage.

Requirements for storage are continually being redefined to meet the enterprise requirements. The costs of storage expansion continue to affect information technology (IT) and the business environment. Observations within the information technology industry have been made to show that storage is utilized inefficiently and that rising costs could potentially bankrupt businesses (Toigo, 2007).

Computing capabilities continue to evolve and storage has never been as accessible, abundant, and cost effective as we are experiencing today. There has been a transition from considering storage as a costly investment to storage as a commodity. The shift in the fundamental philosophy of data storage is affecting everyday operations within the enterprise.

Storage costs alone are consuming approximately 12% to 15% of information technology budgets (Tallon, Scannell, 2009). Storage is more than just physical spinning disks; resources and tools are required to maintain and manage the systems. Information technology budgets are being scrutinized at all levels including reduction to the workforce. The racks within the computer room are again expanding to support the storage requirements.

The significance of this study is to determine whether data and its value are understood within the enterprise. Is data that benefits the enterprise truly being stored or is the problem

deeper, defining a storage expansion model that should not exist? Acquiring an understanding of the data will lead to defining best practices for storage management.

Organizations have become good at the creation of, acquiring and storing data (Detlor, 2010). The enterprise may not fully understand the additional hidden costs that are associated with storage growth and the burden which ultimately affects their revenue. The enterprise continues to lack policies or define best practices for data storage; rather it remains easier to expand than control.

The information technology industry has attempted to regain space by implementing storage recovery solutions. Data de-duplication has become a popular method of addressing multiple instances of a file, but not obsolete, inactive or data that has no value. Lifecycle data management solutions are another method which states that all data does not have the same value and attempts to manage the collected data. The industry has often taken a migration approach to data but the approach only shifts where the data is stored and the cost burden resides. It is unclear whether the enterprise understands the data or its value. Fiscal assets such as data must be defined and understood within the enterprise. Aiken, Allen, Parker and Mattia make a good argument that data management should be viewed as a means to the end, involving both process and policy (Aiken, Allen, Parker, Mattia, 2007). In order to implement policies that can regain storage, a clear understanding of the enterprise's mission, stakeholders and data value is a requirement. The inefficient and ineffective use of data continues to drive the demand for storage expansion; defining policies using best practices will control storage growth.

Research Questions

The intent of the study is to determine best practices that can be defined to control storage growth. This researcher's objective is to answer the following questions:

1. Does the enterprise understand their storage?
2. Is it possible to define and implement best practice policies?
3. Can storage be maximized by assessing the data to verify it meets the defined mission of the organization?
4. How is best practice policies defined?

The objective of the questions is to determine whether data storage growth could be contained if data is understood and valued. The hypothesis is that policies whether initially defined or during a review, fail to consider data value and consider the enterprise mission. Once the policies are implemented they remain static and automated. Enterprise requirements are continually changing and storage should not be eliminated from this equation. If we understand whether policies are reviewed and the data is understood, best practice policies that control or even regain storage can be defined and implemented. The intent of this study is to determine whether the enterprise understands the data it is collecting and outline a process to define best practice policies.

The Value of the Study

Previous lifecycle methods have been define and implemented, to manage data within the enterprise. The focus has been on solutions such as tiered storage or even data compression but not the control of storage growth. Determining whether the enterprise understands their data will define whether information growth can be contained using policies. Understanding the data is essential to managing storage; the enterprise cannot just continue to expand storage. At some point the cost of space, resources and power will dictate that we have reached a storage limit and decision will have to be made that could negatively affect all operations. If processes can be defined that provides direction for storage best practice, growth could be contained and the detrimental effects upon the enterprise avoided. The value of this study lies in the possibility that

best practices can be defined that will provide direction to multiple levels of management, not just information technology.

Result Interest

The information that is provided from this study will clearly be of interest to the information technology teams that exist within the enterprise. But consider further how the enterprise will benefit from this study. In general management within the enterprise will benefit from this study. In fact, providing best practice may not be able to exist without their participation at some level. If management, at multiple levels, understands the goals and intent of best practice storage techniques, the entire enterprise will benefit. Ultimately, the intent of the study is to provide guidelines or visibility into where changes could be made to effectively manage the data and help management.

If best practices to define policies can be implemented and established, management will be able define storage rules that will reduce cost and resource requirements. Any enterprise that has storage growth challenges will want to review and consider the result of this study. This study will provide visibility into possible problem areas within the enterprise that can be addressed directly affecting storage growth.

Chapter 2 - Literature Review

The business model that drives the enterprise has become good stewards of data collection. Data collection has become a defined task expected to produce the enterprise value. Often the tasks have become automated losing visibility into why the specified data is being collected or maintained. Enterprise decisions are not based upon how much data is collected, but data that can provide value to the enterprise. It is irrelevant whether the data acted upon is automated or requires human intervention; the fact remains poor data causes wrong decisions. More data is not better data. The capture of data is driving the growth of storage year after year. It has been suggested that digital data in the world will double every eighteen months (Information Management, 2010).

With the digital data explosion and compliancy requirements, the enterprise has to manage their data appropriately to control their storage growth. Data with value to the enterprise or required for compliancy should be maintained accordingly. The Information Management Journal published that 31% of respondents from a Bridgehead software survey only needed to retain data for four years or less (Information Management Journal, 2007). Additionally, 37% stated they must retain their data in excess of ten years (Information Management Journal, 2007). Clearly this is dependent upon the industry and the size of the enterprise. I found it interesting that the respondent percentages are fairly even. It is difficult for me to believe, that any organization only maintains data for four years especially with today's laws. The survey does however, touch upon a valid point; retention periods have to be defined.

Considerations of data shelf life and value to the enterprise have to be defined and re-evaluated. The growth and management of data has been addressed within the information technology industry, using a number of methodologies. Data migrations, hierarchical storage

management (HMS), data lifecycle management (DLM), information lifecycle management (ILM), and master data management (MDM) are focused upon data and/or information management.

Data Migration

Data migration is a solution that addresses obsolete hardware or provides a mechanism for storage expansion. As hardware fails and spinning disk continues to drop in price, data migration techniques and tools are used to relocate data. Indirectly, it could be stated that data migration can control storage growth. This is a result of data being moved from primary storage to a different, more cost effective storage tier. In simplest terms, a migration is data movement from point 'A' to point 'B'. Space is often not regained but expanded to allow for the continued data growth. Data migrations can, however regain, storage growth if the data is properly evaluated using tools prior to the migration process.

This is not a simple task, since the data resides in numerous data forms such as documents and databases. This type of data is unstructured but clearly may have value. Structured data, such as databases, is much easier to move and evaluate. In either case, multiple instances of a file may reside within storage. An effective data migration must analyze and filter this redundancy of data prior to movement. In these cases where duplicated files are discarded, and data that has value is migrated, storage space is regained. I have been involved with data migrations in the past and from my experience most migrations maintained all data. The enterprise size varied as did their markets. The other clear issue that I experienced was that the entities within the enterprise often maintained their own storage.

Data migrations are a necessity at some defined time within an enterprise. Hardware and software solutions become obsolete requiring new technology to be implemented. Data migration

is however, not a means to an end when it comes to data management and controlling storage space growth. Data migration within most organization occurs for the simple fact that storage growth is required or hardware and software become obsolete. Most data migrations expand storage space even when obsolete hardware is replaced.

Consider further, that the definition of data migration has morphed from the simple to the complex. A good example is the use of policies to migrate data real time. Techniques such as reinforcement learning (RL) algorithms have been investigated and implemented to automatically define policies that can minimize response time (Vengerov, 2008). Reinforcement learning monitors file access and determines the tier where the file should reside.

Policy manipulation is based upon the decision from RL algorithm which are based on statistical estimations (Vengerov, 2008). This solution minimizes the access time risk from the application to the data. The data accessed the most, is maintained on the tier one storage while other data is moved to tier two or three. Storage growth is again contained by moving data between tiers. It is evident that data is being valued and moved based upon access using the RL algorithm. The use of the algorithm provides visibility to data access per unit of time, the requested amount from the tier and the waiting time for the requests (Vengerov, 2008).

Additional research and experimentation has occurred in data migration focusing on access time. A method of migration known as the Aqueduct system, provides a method of data migration during primary hours versus the traditional off hours. Applications can continue to access data in the foreground during the migration process (Lu, Alvarez, Wilkes, 2002). All data remains online. Experiments have been run showing the efficiencies of the Aqueduct system using two baselines: whole store and sub-store. This system is efficient and provides a benefit to the user during data migration; time and data access are addressed not storage growth.

In general, the focus area of the data migration is the time of data transfer. Data compression, hashing and approximation algorithms have been used to maximize data transfer, during the migration process. I believe that the enterprise has become efficient at collecting, moving, storing and maximizing the transfer of data and information. I also believe that data migration could be seen as a promoter of storage growth. Better tools make it easier to continually expand and move the data. What is often not considered is the resources and associated cost. Furthermore, we should consider the data that is being moved. Is the data useful, usable and accessible in the future?

Hierarchical Storage Management

HSM manages data typically using three defined storage tiers. The storage tiers are defined as on-line, near-line and off-line. High level policies are defined that manage the data. The policies simply define the parameters of where the data is to reside and may evaluate file, data and sizes. If the data is moved to a different tier, a pointer record or stub file defines where it resides. HSM became very popular in the early 2000's and continues to be used today. Although the policies are not always complicated, it could be stated that HSM at the on-line and near-line storage tier can control storage growth. It is a superficial solution as the data is only moved not managed. Higher level policies can be effective in controlling storage growth. HSM does not define value of the data being stored but simply manages it through the storage tiers regaining space by eventually moving the data to an off-line or near-line solution. HSM could be considered a simple lifecycle management solution or data migration solution. As for lifecycle, HSM does take data offline but in most instances does not have policies that deletes or eliminates the data. I believe that HSM is more of a data migration solution, based on simple policies.

Lifecycle Management of Data and Information

Storage growth begins with the collection of data. Data collection and its lifecycles have to be defined and supported in order to be effective. There are three commonly used lifecycles that reside within information technology: DLM, ILM and MDM which addresses data at a higher level than those previously discussed. The three methods overlap in a number of areas but their purpose is to manage the collected data at numerous levels within the enterprise.

Data Lifecycle Management

DLM is a lifecycle that is commonly used to manage data. The data lifecycles refers to the capture, sustaining, archiving and deletion of the data when it has met a defined life expectancy. This method is focused upon managing storage at the data level. DLM is the starting point of data collection. It is unfortunate that studies on DLM implementations and solutions have not been published. DLM expands upon the concepts of HSM and data migration within its architecture. From my point of view it is the next evolutionary method to address data and its management.

DLM provides higher level policies that manage the data. Instead of moving the data to offline storage, its policies can delete the data. It is my view that DLM policies must support the mission of the enterprise. I also believe that the data collected needs to be defined by those who are using it. It is understood that DLM can manage data within the enterprise but whether it is the right data is often unknown. This clearly shows that data growth being promoted.

ING Investments has implemented a DLM solution as part of its data management and have achieved a cost savings of 2.31 million dollars (Horwitt, 2008). The success of DLM resides in the defined policies. ING Investments properly implemented DLM to support their

enterprise objectives. If the data being collected and understood at the business level and defined within the DLM policies, then it is an effective approach.

Unfortunately DLM is not always defined as its own solution. It is often merged with an ILM or a part of the MDM solution. When this occurs, the same decision makers are often defining the policies across a number of business and technology areas. DLM then can become ineffective at controlling storage growth. The most essential point is that you have to define data value and be willing to discard anything that does not support the objectives of the enterprise.

Information Lifecycle Management

ILM expands upon the DLM lifecycle. It is again the next step in policy progression and complexity. ILM has to be 100 percent aligned with the business and the business processes. Compliancy, laws and business requirements have to be understood and the information correctly managed. ILM incorporates not only information storage but the recovery, replication, security and content identification. The complexity resides in the policies that interconnect the enterprise. ILM has to drive the decision of data capture. The success of ILM resides in audits, information retention policies, chain of custody, backup and disaster recovery and customization (Socci, 2004). ILM considers what the information is, where it is located, and the relationship to the other information being collected (Petrocelli, 2006).

Additionally, the defined policies do not only consider the storage tiers but how long the information is maintained, its shelf life. The information shelf life is dependent upon a number of factors such as industry and regulations. You would like to believe that there is a direct correlation to shelf life and regulations but this is not always true. ILM can help manage the storage by effectively managing the storage tiers and data shelf life. It is evident that ILM adds a new layer of complexity when it comes to information management; we now have to understand

the enterprise at a higher, more complex level. This is evident with the layer of data protection that ILM implements. There is likely a minimum of two copies of the data being maintained which again increases the storage demand, policy complications and resource requirements.

Ultimately, ILM should discard obsolete or valueless data. Unfortunately not all ILM solutions provide consider the shelf life of data, or looks at the deletion of data as a primary concern. Often access to the data remains the primary goal and concern, even as the ILM storage requirements continue to grow. Petrocelli made a comment that even though storage is part of the ILM picture it is not the complete ILM solution (Petrocelli, 2006). He continued by stating that it is secondary to the process. Implementing a lifecycle does not mean success. I believe that any successful implementation of a lifecycle methodology begins with management support.

Master Data Management

MDM uses DLM and ILM to build a solution that maintains the correct data for the correct period of time. MDM aligns the key stakeholders, participants, and business clients into the business applications, information management methods and data collection (Loshin, 2009). MDM is focused on defining the business rules used to identify and manage data and information. MDM expands across the entire organization and all functional areas. Metadata and data context becomes essential to the successful implementation of MDM.

David Loshin, author of Master Data Management, understands that data management is not a simple task. It continues to be easier to expand than manage the data. Siloes of data often exist throughout organizations with no visibility into redundancy or its value. Vertical lines of business applications have only helped in silo creation (Loshin, 2009). Entities in the organization often have unique applications capturing their own data. Understanding the data continues to be a point of failure. In simple terms, MDM's focuses on the business while

identifying the required entities to define policies, properly ensuring business validation exists (Loshin, 2009).

MDM approaches storage and data from a strategic business direction. Policies have to maintain business and compliancy governance while ensuring data that has value is collected and maintained. Storage growth begins with data collection and the goal of MDM is to reduce the pressure to store and maintain all data through a partnership with management and information technology.

Ultimately MDM is not about reducing storage growth or data. In this case, good data is a requirement and must be factored into the solution. Simply stated, the goal of MDM is to use the essential data to improve the business and reduce risk of failure (Teachey, 2009). Information technology and business leaders may misinterpret the true intent of MDM. MDM ensures that the proper entities are involved and management support exists. MDM does offer valid tools that can reduce and control storage growth but the primary focus areas are risk mitigation, cost control, and revenue optimization (Teachey, 2009). The goal of MDM is to improve business through proper data management. The business focus is essential in controlling storage growth but this is a secondary concern of MDM.

Data Governance

Data Governance defines the decision making processes and rules that are used to manage data and the permissions of who can take actions on that data. It further defines the rules and how they can be acted upon. Vijay Khatri and Carol Brown recommend five domains for the Data Governance decision making process: data principles, data quality, metadata, data access, data lifecycle (Khatri, Brown, 2010). Khatri and Brown define the storage lifecycle as how data

is inventoried, defined, produced, retained and retired while understanding the compliance issues (Khatri, Brown, 2010).

Data Governance is simply a process that applies accountability for data management. In fact there are a number of approaches such as Total Data Quality Management (TDQM), Total Quality Data Management (TQdM), Information Quality Management (IQM) and Data Quality (DQ). The approach defines what is believed to be the best method to achieve Data Governance. Data Governance, like the other lifecycles, is focused on the strategic goals of the enterprise.

The Data Warehousing Institute's survey in 2006 found that 39% of their respondents were applying some form of Data Governance (Russom, 2006). At the same time, 2006, Gartner Research published a finding that stated less than 10% will succeed at their first attempt of data governance (Smalltree, 2006). While Data Governance could be effective in the control of storage growth, the successful implementation seems difficult and costly. Data Governance is focused on the decision making and relationship processes that affects the data. Data Governance does consider the data lifecycle within the relationship but it is again not the main goal. In order to be successful, Data Governance requires a lifecycle approach. ILM, MDM and DLM methods all provide Data Governance a lifecycle that can be implemented into the solution. Data Governance defines the standards, and relationships within the enterprise to employ the DLM, ILM and MDM methodologies. Data governance defines who can do what to the data, at the defined time period. I believe that Data Governance lacks valuing the data simply because the focus lies in the decision making and relationship processes.

Storage Growth

Properly implemented policies within the storage environment can regain space. HSM, ILM and MDM can be implemented to control storage growth and regain it. Information

technology professionals who manage their storage environments are not implementing best practices to ensure storage growth is effectively contained, fulfilling the business requirement. In many cases, storage expansion is the simplest solution to a difficult problem. I believe that the cost and resources of properly implementing a lifecycle solution often out weights adding new hardware. Storage expansion is simply an easier path to follow, rather than understanding the reasons data is collected and its value.

Consider further that IT budgets are constrained while the amount of data, whether for compliance or business requirements, continues to increase at rates of thirty to fifty percent per year (Tallon, 2010). Even with the cost of data storage decreasing, the infrastructure required to manage new storage is often not considered. Organizations are collecting data in every facet of the business with the hopes or possibilities that it may have some value. The information technology groups have to maintain, manage and apply the policies in order to access, store and mine the data. More data, good or bad, requires additional resources whether it is a service, application or physical presence. Within the enterprise are the right people involved with the data management decision process? The reason I ask that question is for the simple fact someone approves the expenditures for storage growth. I further question whether storage growth is understood by the decision makers at the cost level.

Dr. Maes extends the problem of data collection to outsourcing and standardization using solutions such as enterprise resource planning (ERP) (Maes, 2007). There is no dispute that an ERP solution has merit but the question remains, how is the collected data's value defined and who is making the decisions? Dr. Maes suggest that management is not being represented in the decision making process. I believe that this concept must be extended to the business mission. When outsourcing, we expect the organization that is performing the service to understand the

business. Unfortunately this is not always true because the outsourcing company may not obtain or receive all relevant information.

Employees are often reluctant to provide outsides information and maintain their tacit knowledge. In an economic environment, as the one we are experiencing now, gaining a clear understanding of a process or policy may be a challenge. Wrong decisions can be made, which in turn promotes the capture of additional data that provides the business no value. Additionally, poor quality data that is stored provides no benefit to an enterprise. A study conducted in 2002 determined that poor quality data can cost an enterprise 10 million dollars per year (Eckerson, 2002). If proper policies can be implemented within an organization we have controlled storage growth. This clearly is not what most enterprises have experienced.

In a survey conducted by Gartner research in 2010, 47% of respondents ranked data growth as a primary concern (Mearian, 2010). Furthermore consider that 62% of the respondents stated they were planning to expand their hardware capabilities in their data centers (Mearian, 2010). Out of the 1,004 large enterprises who participated in the survey, almost half were concerned about the growth of data, yet 62% were willing to expand. The survey also showed that the respondents will continue to archive data but only 62% will invest in a data retirement solution (Mearian, 2010).

Further indications that storage growth continues to be a problem is the survey that was conducted by the Enterprise Strategy Group in October and November 2008. The survey revealed that 36% of the respondents were concerned with keeping up with storage growth (Lundell, McKnight, 2009). Additionally, 33% of respondents stated that data reduction technologies would be invested in to reduce their overall storage capacity. Of the same

respondents 27% stated they would invest in tiered storage to reduce costs (Lundell, McKnight, 2009). Tiered storage is not a solution but a more cost effective means of expanding.

Within the Enterprise Storage Group report it was stated that data reduction technologies such as data deduplication could be utilized to slow the data growth rates by removing redundant files (Lundell, McKnight, 2009). It was further stated that tiered storage can help control storage hardware costs. Clearly data will continue to grow but even the evidence from this survey shows that there is a lack of understanding in regards to data value. Data deduplication is not the correct solution to minimize storage growth; it is only a temporary cover up of the underlying problem.

Business and information technology are not aligned in regards to the data that is required to support the enterprise. The weak partnership between business and information technology affects the enterprise in several areas when discussing data storage. The most evident is that the entities within the enterprise collect and analyze its own data. The collection of data is not only likely being duplicated, but often data silos exist. Another issue is that the enterprise resources and budgets are being consumed to expand and manage the data. The last concern is whether the data is valid and analyzed properly. Although some may state that the last two concerns are not directly linked to storage growth, the defined storage policies must ensure that they can be addressed. If you are maintaining data that is invalid, you are not only promoting growth but the improper use of resources and poor decisions within the enterprise.

Research conducted by Ventana Research shows that only nine percent of their respondents were satisfied with the quality of their enterprise data (Ventana Research, 2009). Additionally their research has shown that 59% of their respondents reported disparate sets of data-handling and collaboration technology to be their top barrier (Ventana Research, 2009). It is

evident that storage growth is being promoted from a number of forces within the organization. Data, whether structured or unstructured is not the major factor promoting storage growth within the enterprise; it is the understanding of the data.

Policies and Data Management

The management of data has become a staple of business success. The enterprise relies upon policies to manage, consolidate, control access and store the data within the enterprise. Unfortunately, many organizational policies do not exist to evaluate storage and storage growth. In simple terms, the lack of policies cost the enterprise revenue. Paul Tallon stated that the backup, recovery, labor, and overhead of data storage is often not considered during the data management process (Tallon, 2010).

Data management terms have become common place in the information technology industry. The Information Management Journal states that 62% of people are migrating data due to growth and 58% due to compliancy (Swartz, 2007). Storage growth continues to be a concern. Understanding the relationship between the data, value and business mission promotes proper storage. Detlor has written that three major perspectives have to be considered within data management: the organization, library and personal perspective (Detlor, 2009). Information lifecycles define data value. What lacks is applying best practice policies in a relevant way to constrain storage growth. The IT industry has become focused on data availability and analytics not data value. Businesses continue to drive storage growth but it is not always clear if there is a business requirement. Dr Maes, a leader of the PrimaVera research program, believes that we do not require more data but a better understanding of the data (Maes, 2007).

Ciroth's study further enforces this statement by assessing cost data quality using a pedigree matrix. The basic quality indicators of the data are reliability, completeness, temporal

differences, geographical differences and technological differences (Ciroth, 2008). Although the pedigree matrix in the study is focused on data value for cost, an understanding of what may be required to control storage growth is becoming more evident.

Other solutions continue to be studied which would manage massive amounts of data. Consider the scientific research collaborative and the data they generate. Agreements are being defined that share data across all parts of the world; we have experienced this within universities. The integrated, rule-oriented data system (iRODS) method approaches the mass of data as a global share that reduces duplication of effort and storage costs. The iRODS approach is a data grid that spans sites supporting a policy based life cycle management (Rajasekar, Moore, Wan, Schroeder, 2009). The iRODS solution is unsuccessful at providing a unified repository of information without policy enforcement. This is not a negative comment against the iRODS solution but verification that defined policies must be enforceable, else they do not provide the enterprise any value.

Data storage growth is directly related to data management and the policies defined by the management. With the explosion of data capture, the constraints on transfer, recovery and backup continue to be challenged. David Loshin states that every organization using data is plagued by poor management practices and it is easier to treat the symptoms than to resolve the problem (Loshin, 2009). David Loshin has recognized that we do not evaluate the collected data but continue to relocate it or expand the storage requirements. In essence we have become our worst enemy when it comes to properly managing data collection and storage.

Policies are defined as best practices that support the mission of the business. ISO 14040 simply defines policy as data characteristics that have bearing on the ability to satisfy stated requirements (Ciroth, 2008). Others have defined a policy as a set of best practices that an

enterprise has to follow (Petrocelli, 2006). Petrocelli further expands on the definition and states that policies are concrete expressions that support data protection strategies (Petrocelli, 2006). It is imperative to understand that policies exist at both the human and technology level. Petrocelli makes the point that policies may not value data properly and in fact, value all data the same. If policies treat data of little value at the same level as data that is critical, then valuable resources are utilized for the wrong purpose. Although in this case, he was not referring to storage, it does apply. Storage is a resource that policies need to manage. Data with little or no value must be dealt with differently than critical data. How the policy is defined affects all storage tiers.

Policies are often defined by using a top-down or bottom-up approach. Top down is often the most popular approach for data management; upper management support is required. Petrocelli stated that top down requires significant work before it can become useful (Petrocelli, 2006). The bottom-up approach is possibly more relevant when it comes to data storage because it reflects the value of the data based on the users. The bottom up approach focuses on the functional areas within the enterprise allowing detailed policies to be developed quickly which can be put into action sooner (Petrocelli, 2006). Petrocelli believes that the top down approach works the best. From the perspective of storage growth the bottom-up approach seems to provide a better method to define use and value, but lacks upper management support which inhibits success. Success requires support from upper management and the approach has to ensure it is present prior to implementing policies.

In general, all research agrees that successful policies must be written down, enforceable and define the intent or rules clearly while supporting the mission of the organization. Data storage policies have to define data value and quality. How many policies within an organization enforce data and information management? Who is making the policies that decide the data

management processes? Often the information technology management is making the policy decisions. These strategies of policy decision making may not align with the enterprise. Storage growth may increase at unprecedented rate because defined value is lacking.

The Data Management Body of Knowledge (DAMA-DMBOK) recommends the use of data stewards. Data stewards have the best interest of the enterprise, management and information technology in the design of the policies. Their job is to understand the enterprise requirement, and ensure that the data being captured by the defined policies supports the mission. If policies are defined and implemented properly, redundancy is contained. The DAMA-DMBOK has documented the following rules to contain data redundancy: govern the creation, acquisition, integrity, security and use of the data (Mosley, Brackett, Earley, Henderson, 2009).

Depending on the enterprise and the storage implementation, redundancy of data could be significant. The rules above support the regain of storage space. Redundant data provides no value especially if data is being captured by different entities at different storage locations. Information technology professionals strive to balance storage against a number of requirements within an organization. There are a number of constraints that have to be addressed to regain storage such as tools, resources and management support. To meet these objectives, policies should be defined that encompass an understanding of the stakeholders, the enterprise, data types and compliancy requirements. David Loshin makes the point that operational and analytical processes must clearly define the business concepts (Loshin, 2009). In addition the direct and indirect benefits, cost of loss, and impact to the enterprise must be considered (Mosley, Brackett, Earley, Henderson, 2009).

Policies that control data and the growth of it are not easily implementable. Information technology professionals have to continually evaluate the storage environment to ensure that they

can support the day-to-day requirements. A data reform program that was implemented in an Australian healthcare environment substantiates the need for policies that support data understanding. The problem was approached by reducing the proliferation of program data collection and improved data integrity to increase the impact of the data (McKenzie, Perry, Ashley, Dalton, 2010). Data storage growth is not only driven by our ability to capture more data but the inability to understand the data and define polices to address it properly. The challenge for the enterprise continues to be defining the value and shelf life of the data. In many solutions context and metadata are relied upon for policy support.

Context and Metadata

There have been numerous articles and books written on data context and metadata. Data value changes continually in the enterprise in which directly affects storage growth. Simply put, new valuable data in, obsolete data out. Eckerson stated that two percent of data becomes obsolete in a customer's files in one month (Eckerson, 2002). Although the customer data is affected by a number of factors, all data collected has value for a defined period. Data collected from websites for product analytics is normally based upon human interactions. The data is only good until the next product is released. Data's value resides in building the relationships to solve a problem or provide information to the enterprise. Data that cannot be easily found or evaluated only expands storage and provides the enterprise no value. Data that does not hold value should be deleted.

The DAMA-DMBOK states that without context, data is meaningless (Mosley, Brackett, Earley, Henderson, 2009). The Data Management Association has identified the context around which data should be meaningful. Meaningful data has (1) business meaning of the data elements and related terms, (2) the format of the data is present, (3) the time frame represented by the data

and (4) the relevance of the data for the given usage (Mosley, Brackett, Earley, Henderson, 2009). David Loshin enforces the concept by stating that data sets must solve a business problem (Loshin, 2009). If we understand the business problem that is being solved, the collected data can be understood allowing policies to be designed either to maintain or discard the data. Applying an interpretivist approach to data that has meaning, provides a method to regain storage and define best practice.

Petrocelli also believes that context is critical for the success of data management. He defines context as a perceived set of variables that may be of interest to the enterprise which can influence the action or decision (Petrocelli, 2006). He further states that there are two types of context: explicit and implicit (Petrocelli, 2006). Explicit context is easily understandable like that of a database. Documents would be considered implicit. Explicit context is much easier to evaluate and resole than implicit context. Context provides the additional data ensuring that a clear and concise understanding of the primary data exists. Context is an important tool in the enterprise. If context is utilized properly it can be used to define data value. Data that is not understood is not of any value and continues to expand the storage requirements.

In order for context to be effective, it is clear that an understanding of the business is required. In regards to storage growth, it is essential to understand the data across the enterprise which includes storage, time and space. Carter and Green make a good point that information is contextualized data (Carter, Green, 2009). They recommend three focus areas: Country specific, firm specific and key management / information technology issues.

The country specific areas refer to compliancy laws, regulatory factors and political concerns. Political concerns can be related to external forces. Firm specific refers to the strategies of the enterprise. IT management issues are concerned with the internal forces and

politics of the business. For context to be effective an open and complete view of the data must exist. Carter and Green state a holistic view is required to be successful. The Homeland Security initiative is a good example of expanding data that in essence fails to create usable information. Data that has value but cannot be manipulated into usable information provides, the enterprise no value. Simply put, storage resources are used to maintain and evaluate data that is unusable. Collecting the right data does not mean it is usable within the enterprise.

Green and Carter have noted three key problems with Homeland Security data collection policies: data existing on multiple platforms, a lack of data quality, and inadequate metadata (Carter, Green, 2009). The concerns that Green and Carter have stated are relevant in any enterprise that is dealing with the management of data. Green and Carter use an illustrative case to show how context could improve data quality while reducing storage requirements. One of the major problems as seen across numerous enterprises is the sharing or use of data. Because a mix of disciplines existed, data management tasks were being performed independently. This only promotes silos of storage, increasing data duplication, additional resources and storage expansion. In this case scientist had the decision making authority.

Context is still dependent upon a number of factors. Communication, coordination, monitoring mechanisms and training must be present for success (Carter, Green, 2009). Secondly, from the above case, it is clear that data reuse was not occurring across the agencies. Storage growth continues to be promoted rather than addressed properly. There are a number of implemented solutions known as context-aware systems. Context aware systems rely upon context to make decision on how to handle the data. Solutions such as ACTIVITY, CASS, CoBrA and COMMANTO rely on context.

More often than context, we hear the word metadata. Metadata in many ways seems to be the same as context. Most define metadata, as data about data. There is a clear relationship between metadata and context. Context defines what the metadata will provide. If context does not exist the metadata will not be understood. Metadata provides meaning to the data. Metadata has become an essential tool for file management.

Metadata clearly reduces search time for information and promotes efficiencies within the enterprise. If metadata is defined properly it is easy to reuse the existing data and again minimize data duplication. Considering business analysts spend approximately 30% of their time analyzing and validating collected data, metadata provides the enterprise a benefit (Inmon, O'Neil, Fryman, 2008).

Although metadata is very promising it contains challenges. The largest challenge to overcome is the relationship between context and the metadata. It is clear that context and metadata require the right involvement within an organization in order to maximize the full benefits. Further, it is clear that multiple forms of metadata exist. For instance, Microsoft Word creates a level of metadata automatically when a document is created. Because this is unstructured data, it is difficult to say it will solve the problem. Agreement within the enterprise and the ability to capture the same metadata throughout all applications is a challenge.

Microsoft Corporation performed a five year study using metadata within their organization and the Windows operating system. Their participation rate for the five years was 22% of the employees (Agrawal, Bolosky, Douceur, Lorch, 2007). The study collected data on a number of areas: file size, type frequency, storage capacity and storage consumption. The study was not geared at controlling storage, but did allow Microsoft storage planners to utilize the data.

Files that are modified locally, typically refers to multiple copies that reside within the storage network. The study found that in year one the median-file system had 30% of its content modified locally. In year four, local file modification was reduced to 22% (Agrawal, Bolosky, Douceur, Lorch, 2007). It is feasible that storage moved from predominantly local to network storage causing the file modification location to change and the percentage to look better. If this is the case modification was still occurring but at a different location.

From the use of metadata, Microsoft found that file system storage increased from seven percent to fifteen percent. Additionally, they determined that the files system media capacity rose from five gigabyte to forty gigabyte (Agrawal, Bolosky, Douceur, Lorch, 2007). The study consisted of 63,398 distinct file systems (Agrawal, Bolosky, Douceur, Lorch, 2007). Microsoft's study provided a good view of how metadata can provide significant information on files size and growth but this may not be realistic in the enterprise. Microsoft's advantages were the use of the same operating and file system throughout the enterprise. What Microsoft proved was that metadata can answer questions and provide insight into file changes and growth which provides the enterprise valuable information that could be used to control a number of factors within the file system.

Conclusion

A number of lifecycle methodologies and even storage solutions such as HSM are implemented and used within the enterprise. Journal articles, books and solutions have been defined to solve the storage problems. Defined implementation versus practical implementation in many cases is two different solutions. Weber, Otto and Ősterle stated that research on data governance is still in its infancy (Weber, Otto, Ősterle, 2009).

The fact remains that storage growth continues to expand at significant rates. Simply put, we collect more data not because we have to but because we can. Data has become a valuable asset to the organization and we have become data hoarders. One possible reason that data hoarding occurs is because at some unknown point data loses value but there is no method to determine its changing value. Another possible reason is that the continual drop of magnetic disk storage provides a simple means of expansion. The question remains, where does the problem lie within the enterprise and what are the best practices to control storage growth? Are lifecycle policies the best method of storage growth control and does metadata solve the valuing problem of the data? Storage growth problems continue to be an industry concern. When 47% of respondents replied to a Gartner Research survey stating that data growth is one of their top three concerns, we have not addressed the problem (Mearian, 2010).

If we recap, the lifecycles that are offered to the enterprise are built upon and supported by policies. It appears that the policies primary objective within the enterprise is to provide access to the data and migrate it to more cost effective storage tiers. Data access time remains a key component defining storage management success. The tools exist, but policies are not defined to control storage growth. Even if the right policies are implemented is it possible to maintain them?

From the literature review, there are a number of data management lifecycle methodologies offered to the enterprise. In my opinion, many of the methods defined do not have unique or different approaches. Yes, they may use different headings but often the intent remains the same. Does the enterprise understand what the solutions industry is selling and solving? Further, does the enterprise's information technology staff understand their storage policies and how it supports the mission of the organization? If the mission of the enterprise is not

understood, can a methodology be applied from a solution provider to address storage growth?

In conclusion, what are the best practices that can be applied to control storage growth? Do the lifecycle methodologies overlooking the obvious within the enterprise that promote solution failure? I believe that the issues are deeper than the lifecycles that have been discussed within the literature review and begin prior to any solution implementation. Furthermore, I am not convinced that data collection is understood at the enterprise level.

Chapter 3 - Methodology

Defining best practices regarding storage growth requires a clear understanding of the problem. To acquire an understanding, research was performed that supported and substantiates what I believed to be factual. In order to collect the data, a mixed method approach was taken. The initial data was collected by administering a survey. In addition to the online survey, two industry professionals were interviewed to further substantiate the survey findings.

Participants

Survey

The survey consisted of twenty eight questions which took no more than ten minutes to complete by the subjects. I had two primary objectives when administering the survey: maintain subject anonymity and collect data from participants who had experience with storage, data migrations and/or archiving. The intention of the survey was to encourage subject's participation which included not only information technology professionals but management as well. This would provide a broader cross section of the collected data. The target for the survey was storage experts, information technology management, information technology administrators, business management, chief information officers, chief executive officers and system engineers. The first survey question asked the participants to choose the role that best fit their job description. Table 1 provides the response according to profession. Table 1 depicts that information technology management had the highest response rate for the survey.

Table 1: Subject Participation by Profession

	Quantity	Percentage
Storage Expert	4	15.38%
IT Administrator	5	19.23%
IT Management	8	30.77%
Business Management	4	15.38%
CIO	0	0.00%
CEO	0	0.00%
Systems Engineer	1	3.85%
Other	4	15.38%
Total	**26**	**99.99%**

Interviews

Two interviews were conducted to further substantiate the survey findings. The first interview was conducted with an ILM industry expert. This individual has written a number of industry articles and conference speaker regarding ILM implementation in the medical industry. This individual has industry experience and has worked for a number of large technology companies. The interviewee is currently working to educate medical storage vendors and customers on the value of data. I was referred to this person by an industry storage expert.

The second interview was conducted with an individual that I had worked with previously and who has started a data migration company. When I began considering my thesis topic, this individual allowed me to utilize newly developed software to review filtering techniques that were being incorporated into the software. Interviewee two understands that all data does not require migration and should be able to have a filtering mechanism within the process. This individual has worked for a number of large technology companies and has a very well rounded understanding of data storage. The individuals will be referred to as interviewee one and interviewee two to maintain their anonymity.

Admission

Survey

Prior to any participation in the survey, consent was required. The survey was preceded by the approved consent form and was accepted by pressing the agree button at the end of the form. If the button was not selected the survey exited. The consent form stated that the collected data would be used to understand storage growth and did not collect any participant's personal or organizational information maintaining their animosity.

The survey was distributed using a professional network. Groups that fit the goals and focused on storage were targeted. Professionals that worked in the storage profession were targeted. Furthermore, the focus groups dealt with storage, data migration, and data archiving. The group was notified of the survey using a message post within the focus group. Each focus group area had the same message published with a link to the survey. This ensured that I would not know the participants and they would remain anonymous.

Interview

The interview subjects were sent consent forms prior to the interview. The two interviews were scheduled based upon a mutually date and interview location. The interviews were scheduled for one to two hours and could be recorded with their permission. One of the two interviews was recorded. The subjects were aware of the topic from previous communications and were required to sign the approved consent forms prior to commencing with the interviews.

Research Materials

Survey

The primary instrument that was utilized for the research study was a survey. The survey consisted of 28 questions. Each question consisted of three to seven multiple choice answers.

The average time to complete the survey was ten minutes. The questions were focused upon determining whether the enterprise understood their data and storage techniques. The goal of the questions was to determine whether storage expansion was occurring and the level of data valuing. It was not the intent of the survey to review policies that control storage but to understand the decision making process used to define the policies that exist within the enterprise today. This method allows best practices to be defined that address the storage growth problem.

The 28 questions were based on multiple choice answers that addressed multiple views, providing the subject the ability to choose the best answer. Some questions allowed for "other" to be selected and a unique answer input. Percentages or ranges were used in multiple survey questions to build a better understanding of where the issues resided for the majority of the participants. The survey questions can be located in Appendix A.

Interview

The interviews were used to substantiate what was found in the survey. Questions were formulated to provide a deeper understanding of the real world problem outside of the literature and survey. The intent was to expand my understanding of the problem and what has been tried within the enterprise. The first interview was based solely upon the medical industry while the second covered multiple industries. The interviews were conducted at a restaurant that was mutually acceptable. Both interviewees' discussions focused upon their area of expertise.

Procedure

I had to determine how to effectively and efficiently distribute the survey. I believed that the most effective method was to use an online survey hosting solution. After research of my available and/or affordable options, an online hosting service was chosen. On November 20, 2010 the survey was posted. In consideration of the Thanksgiving holiday I did not make the

survey available until December 3, 2010. The intent was to keep the survey active for 20 days and review the responses. Initial response was limited which I felt was due to the holidays. I chose to extend the survey for an additional three weeks. As shown in figure 1 the response and completion percentage was respectable once the survey was closed. The survey only allowed one subject to participate per access point.

Once the survey data was collected it was downloaded into a spreadsheet for further analysis. Cross tabulation was available using the tools from the online host. After the survey data was reviewed, two interviews were scheduled over a two week period. The interviews were recorded for further analysis and further questions.

Data Analysis

The first step that I took in data analysis was verification that I had enough respondents who completed the survey. Figure 1 shows that 79 viewed the survey and 26 completed it.

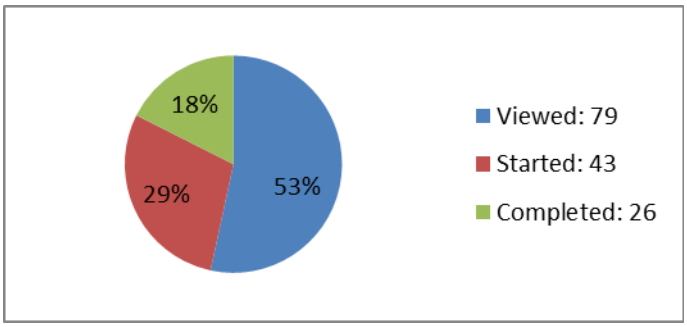

Figure 1: Survey Participation Statistics

The 43 surveys that the subjects did not complete were not counted in the data analysis. The average standard error of the survey questions was .1977. Trends were formed from the survey showing the subjects response to the storage questions. The data was further analyzed by reviewing all of the answers to gain an understanding of the problems. For instance, Table 1 showed that approximately 31% of the respondents are information technology management but

understanding their company size may allow for better analysis of the data. I also wanted to see diversity in the respondent's profession so that multiple views were captured by the survey.

The percentage of a selected answer by the participant, defined what the enterprise was experiencing. Using an interpretivist approach to the survey answers allows the problem areas to be determined and best practices to be defined.

The survey questions were designed to support the four defined questions in the thesis introduction. A single survey question does not fulfill the requirements to make a decision, it takes many. The first thesis question, "Does the enterprise understand their storage" relies on a number of questions to formulate a conclusion. Twelve of the survey questions provide some level of information to answer this question and some interpretation is required. A good example is question five that queries whether the stakeholders within the enterprise are involved with the storage decisions. Asking this survey question defines whether the right people are involved with the storage decisions.

The enterprise is run by management and their decisions drive the storage requirements but do they really understand the cost of storing the data? The survey addressed this with the participants. Additionally, other survey questions asked the participants whether data is valued or storage is expanded. Evaluation of multiple survey questions will provide an answer to the first thesis question.

Does the enterprise rely upon the entire organization, stakeholders, and business management or information technology groups to make the decisions? If they rely upon only information technology do they really understand their storage? The question was further supported by the conducted interviews. Interviewee one and two presented power points to me substantiated that storage will continue to grow and that the collected data is not understood.

The second thesis question considers whether best practice policies exist to manage storage growth. Six other survey questions indirectly address this with the subjects. Again, an interpretivist approach was taken because variances exist within the enterprises. The data analysis from the survey questions continue to expand my knowledge of what is implemented and defined within the enterprise.

I needed to understand whether the defined policies truly manage data or is it simply being moved and stored. Specifically how does the enterprise handle data with little value? Survey question thirteen addresses value with the subjects and allows them to choose the answer that best fits their environment. One survey question does not alone provide an answer; multiple questions are interpreted to build the answer. The two interviews help to substantiate the findings of the survey.

The third thesis question addresses whether storage can be maximized by assessing the data. Eight survey questions help to address this question. The survey provides an understanding of how often storage is expanding within the enterprise and establishes a time frame for the expansion. The survey further support answering this question by asking whether the data value is determined within the enterprise. If data is properly valued and storage growth continues to occur, you can formulate an opinion that either all the data is required to support the business or data is evaluated but not properly utilized.

Other survey questions expand upon the storage growth area building a view into the systemic storage growth problem. If the tools exist within the enterprise to value or assess the data, then storage policies are not aligned with the business. Understanding this alignment is critical to the success of this study. Additional survey questions address whether data evaluation has been established. The lifecycle management question within the survey provides visibility

into whether one or more solutions are implemented today. If an enterprise implements more than one lifecycle, can it be properly managed and what mission does it support? In support of answering question three, understanding data value is essential. If data value is not understood how are policies defined?

The final thesis question addresses how best practices are defined. All of the survey questions help support the resolution. Best practice can only be defined if there is an understanding of what affects the enterprise. The analysis requires a wide view and understanding to sufficiently answer this question. Analysis of the survey question responses has clearly defined that policies are enacted that consider data at some level. Although policies exist, this does not define them as a best practice; it is also the interpretation of the individual taking the survey.

In order to provide a solution to this question, it requires an understanding of the problem from multiple vantage points. The survey cannot implicitly solve this problem but all the questions together provide direction for best practice. Furthermore, the interviews became a key component to my evaluation of the data. The individuals interviewed have firsthand experience where the problems exist. The solution to question four lies in understanding of what the subjects are doing within the enterprise. When I take the discussion information and apply it to the survey then finding an answer can be formulated.

The methodology approach employed in this study is mixed methods. The survey is based on the percentage of answers returned by the subjects. Further investigation was required even with the response from the survey participants. Using both the interviews and survey, I created a complete picture of the problem and how best practices can be defined.

Chapter 4 –Results

The survey results show that storage requirements are growing at incredible rates within the enterprise. Furthermore, the results provided a solid understanding and insight into the issues and where the problems exist. The survey and interviews focused on the topic areas to answer the questions that reside in chapter one.

Storage Growth from the Survey Data

The enterprise captures data in order to run a successful business. Data provides the enterprise the ability to perform analytics, develop business and meet the legal requirements. The survey data shows that the majority of the respondents reported that their storage growth is between 11 percent and 50 percent annually as shown in Figure 2. The survey further revealed that no participant's storage has decreased their storage requirements.

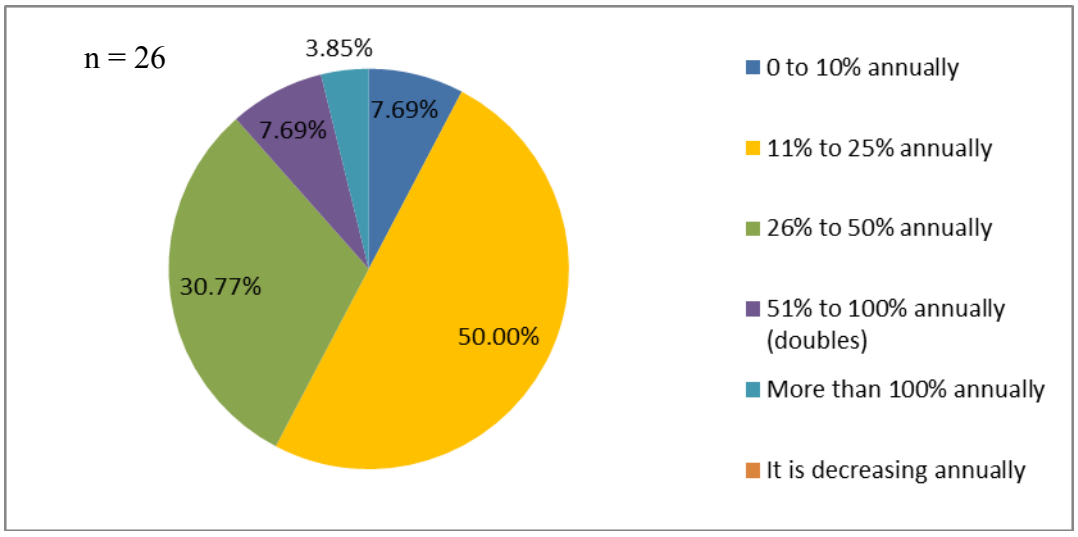

Figure 2: Percentage of Storage Requirements based upon Annual Increase

The survey further allowed for a model to be built that provides an understanding of the effects of storage increase over time using a fixed starting point. Figure 3 is based upon existing storage of 100 terabytes. Figure 3 shows the effects that storage expansion has upon the

enterprise. Using this model I have determined that 31 percent of the subjects will double their storage by year two. Additionally, 69 percent of the subjects will double their data in approximately year three.

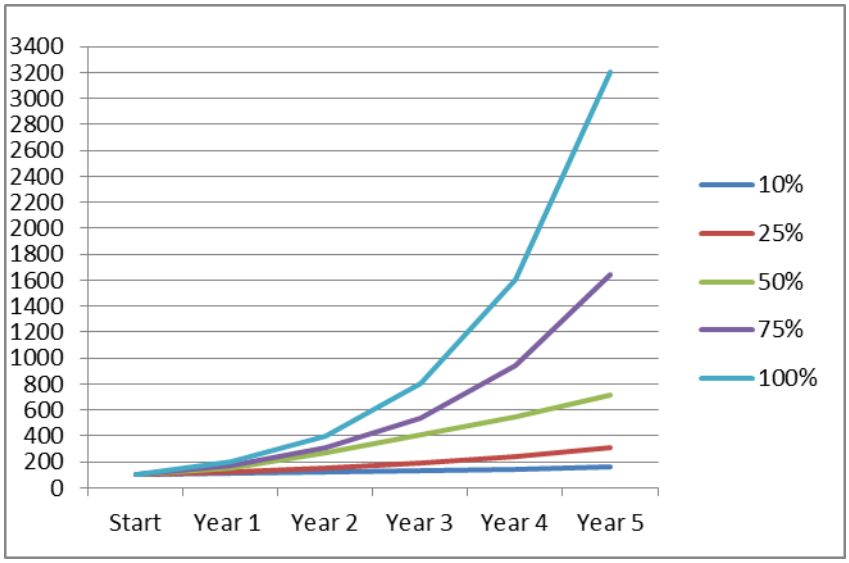

Figure 3: Five Year Storage Growth Projection Based Upon 100 Terabytes at Year One.

Cross tabulation of the storage growth and organization size was used to determine whether the size of the enterprise was greatly affecting storage growth. Aligning storage growth with the enterprise size provides an understanding whether larger organizations manage their data more effectively. Performing this analysis, I determined that 53 percent of the subjects with 250 or more employees stated that their storage growth would expand between 11 percent and 50 percent annually. One additional subject with an organization size of 250 to 500 employee stated their growth would increase 100 percent annually. This was only one data point so the margin of error is higher. The data provides evidence that large and small enterprises alike are similarly increasing their storage requirements.

Implemented Lifecycle Methods

Lifecycle management of data is a common solution that is often implemented in the enterprise. The survey asked the participants two questions regarding to information lifecycle

management (ILM) and data lifecycle management (DLM) implementations. The survey has shown that 53 percent of the participants have implemented a DLM, ILM or both while 35 percent do not have any solution implemented (31 percent + 4 percent).

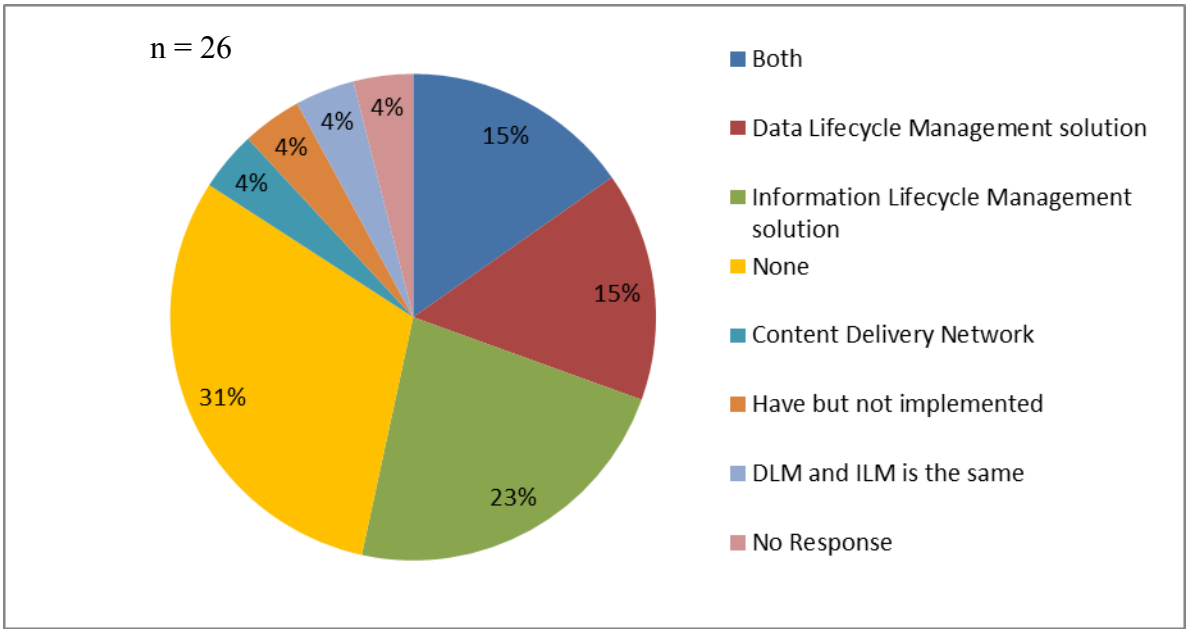

Figure 4: Implemented Enterprise Lifecycle Management Solutions within the Enterprise

Figure 4 shows the percentage of the lifecycle implementations. Furthermore, the data shows that other solutions are being implemented by a small percentage of participants. One participant stated that ILM and DLM are the same. ILM and DLM was discussed with interviewee one. Interviewee one stated that ILM and DLM are not the same; ILM is more complicated and DLM is more of a hierarchal storage management (HSM) solution (personal communication, January 23, 2011). Additional evaluation of the data using cross tabulation determined that 27 percent of the participants who had a lifecycle solution implemented, stated they would need to expand their storage within six months. The data further showed that an additional 26 percent of the participants who have implemented a lifecycle solution would also need to expand their storage within a year.

The survey participants all understand that their data is growing at some specified rate as shown in Figure 2. The survey data further shows (Figure 4) that lifecycle solutions have been implemented in an attempt to manage the expanding data challenge.

The second lifecycle survey question focused on the implementation. The questioned asked whether they managed the data properly. From the data, 32 percent stated they address the data properly while 16 percent stated a conflict existed. Another 4 percent was unsure of the functional status. The survey has shown that over 27 percent of the participants do not use or have some conflict in their storage management solutions.

Additionally, 48 percent of the participants made unique comments in regards to a lifecycle implementation.

- Your questions do not apply to my organization, as you appear to be focused on file management and database storage rather than CDNs, asset management, and video.
- We do not use either of these solutions. Data is not purged today except in a few cases where storage expansion is dramatic for the given table.
- Do not use.
- Not applicable
- Not used
- The conflict is due to lack of review because of low resources. So, data is kept forever.
- Depending on the resource and type of data, there is often some conflict.
- N/A
- Data is migrated to slower online and long-term near-line storage over time.

Storage expansion and data value

Understanding data means that some value is assigned to it. The survey used seven questions to build an understanding of data valuing within the enterprise. Participants were asked whether it was easier to expand storage rather than define or re-define the value within the existing policies. The intent of the question was to define whether the enterprise believes that valuing data is important. The survey data has shown that 20 percent of the participants would review the defined data value prior to expansion. Figure 5 provides a view of how the survey respondents are managing their expanding data. The majority of survey participants, 80 percent, choose to expand rather than define the data value.

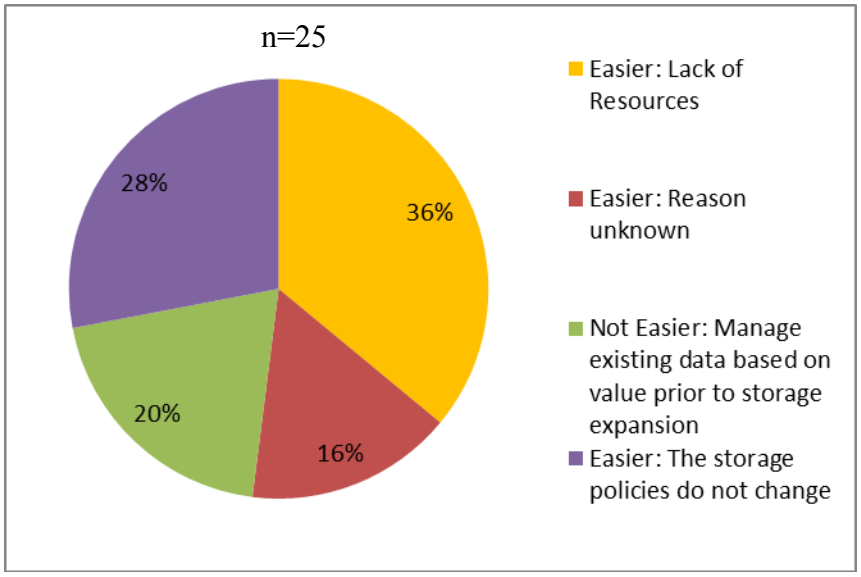

Figure 5: Storage Expansion is Easier within the Enterprise

Using cross tabulation against the survey questions that addressed storage expansion and the implemented lifecycle management, I further determined that 36 percent of the participants who had a DLM, ILM or both lifecycle solutions stated it was easier to expand storage than value their data. The survey further asked that if your storage could not be expanded would you make it a point to understand the data or implement the first-in-first-out (FIFO) approach. The

majority, 88 percent, stated they would value the data while eight percent would use the FIFO approach. Interviewee two stated that politics within the organization drives the information technology staff to over buy and store everything.

Understanding the mission of the enterprise is important to all levels of the business. The question was asked of the survey participants whether the collected data is maintained because it is believed that it has or will have value. The "yes" response was over 95 percent. This response was expected, however I wanted to establish that the enterprise finds it easier to expand storage rather than value or assess the data and that all the stored has some value.

The question was asked to the participants whether the data is maintained because of compliancy or government requirements. The surveyed showed that 69 percent stated that the collected data was maintained to meet compliancy or governmental requirements while 23 percent said that was not the reason. The follow up question asked, whether the value of the data is measureable in revenue, product or future offerings?

The survey showed that 46 percent of participants stated that the value of the data was measurable in revenue, product or future offerings while 35 percent said it had no value and 19 percent were not sure what the data was used for. Simply stated, 54 percent of the respondents do not believe or are unsure that the data they are collecting brings value to the enterprise. The survey data showed that data is being maintained that may not have any value.

The survey presented two questions that were closely aligned in regards to data valuing. Data valuing provides a method to assess the data to verify that it meets the mission of the enterprise. The first question addressed whether or not the enterprise classified its data to delineate which data has the most value. The second question asked whether the data management policies consider the value of the data when it was originally stored. In both

questions over 50 percent stated their enterprise and policies considered the value of the data. Almost equally, 42 percent of the participants stated that they did not value their data at either level.

The answers have provided more than just a percentage of response in regards to the valuing data. The enterprises that value data are doing it at the management and policy level. Interviewee one provided additional insight into data valuing (personal communication, January 23, 2011). In the discussion it became very clear, that the data must be understood to define the value and often the value changes daily; it is dynamic not static.

Interviewee one stated that data is often misunderstood and stored at the wrong tiers or failed to be discarded (personal communication, January 23, 2011). It serves no purpose to continually store data with no or little value (personal communication, January 23, 2011). Interviewee two presented a different view from experiences with data migration. Interviewee two stated that the enterprise often has silos and that everyone who has a silo wants to control the value of the data (personal communication, February 7, 2011).

In essence the premise is that the person with the most power may dictate how data is valued across the enterprise (personal communication, February 7, 2011). An example that was presented was from the discussion with interview two was that the information technology staff reported to legal and legal defined the data value (personal communication, February 7, 2011). Legal only understands data value for laws and compliances but not the other areas of the enterprise such as finance (personal communication, February 7, 2011).

The survey further considered whether management of the enterprise understands the cost of, storing and analysis of the data. If data is being maintained with little to no value it is very likely that application and resources are continuing to perform analytics upon it.

The survey asked the participants whether their enterprise considered the cost of storing, analyzing and moving the data. It is reasonable to believe that within the enterprise cost of storage expansion, data analysis and data movement has been considered. Figure 6 shows the responses from the survey question.

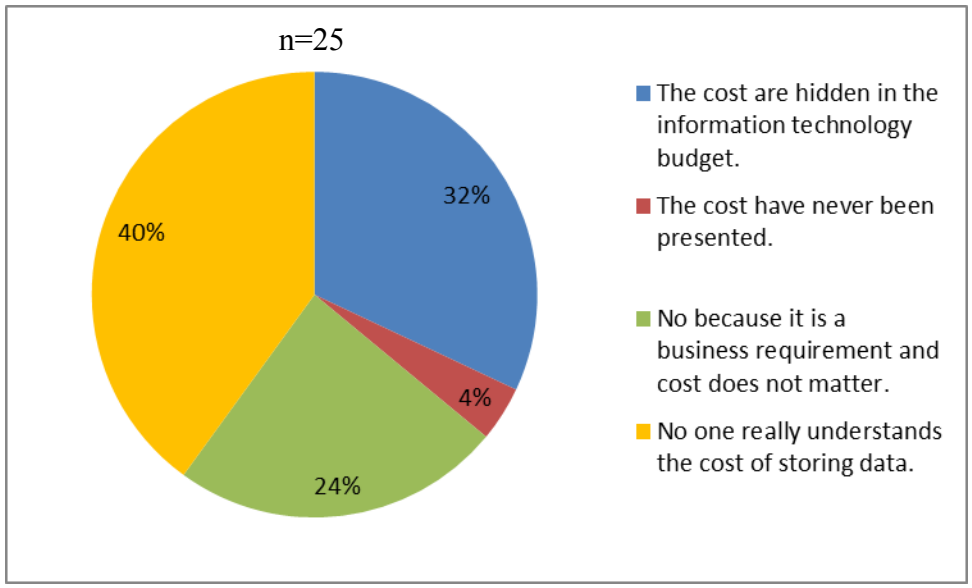

Figure 6: Response to the Cost of Storage, Analyzing and Moving Data

The results clearly show that the cost of storage is not understood as denoted by 40 percent of the participants or hidden within the information technology budget as depicted by 32 percent of the participants. Furthermore, the survey has brought to light the fact that four percent of the participants stated that management has never been presented with the cost of storage.

Management Participation

Management is essential to the business decision process. Without the right data they cannot make knowledgeable decisions. Understanding whether management is involved in the decision making process of valuing the data is important to understanding the problem of storage growth. The survey asked that if their management understood the cost of storing data would their participation change? Figure 7 presents the results.

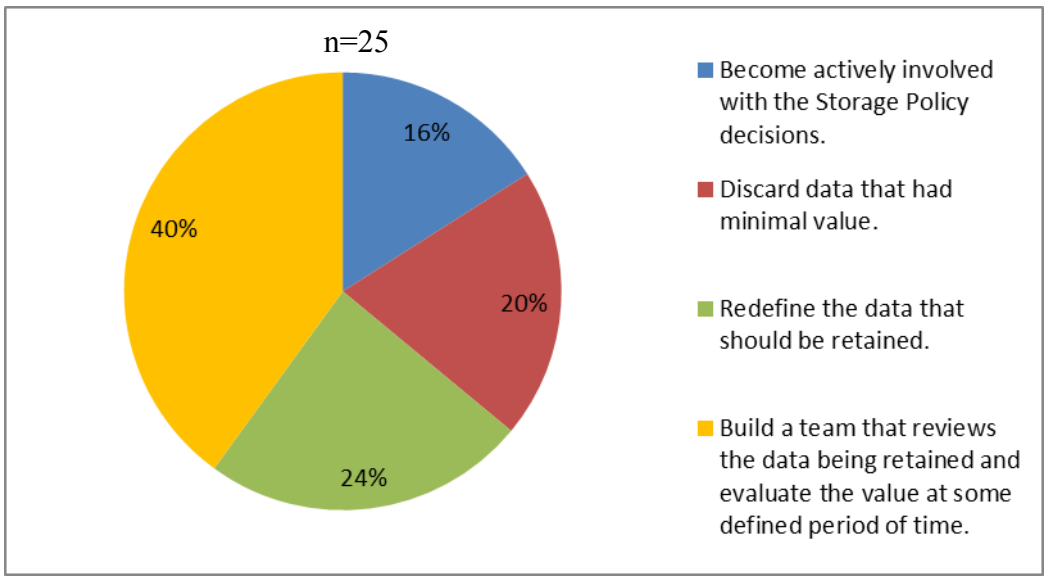

Figure 7: Management Decisions based upon Understanding the Cost of Storage

The participant response showed that 40 percent of the participants would build a team to evaluate the data. Figure 7 provides more than just percentage of response in regards to policies and data value. The data shows how management would react if storage cost were understood. The question further brought to light that understanding storage cost is lacking. Furthermore, there is almost an even split between evaluating the data that should be retained and discarding data with minimal value. The data provides information on how management will address the storage situation in regards to policy implementation. There is a clear difference between data discard that has no value and redefining the data that should be retained.

After analysis of these results I felt it would be good to understand where the participant's population lied in regards to enterprise size. Understanding whether the size of the organization had any bearing upon the decision making process in regards to storage cost, would provide an additional trend for policy definition and understanding of storage. Performing a cross tabulation upon the organization size and management decisions based upon storage cost produced the data in Table 2. The table is based upon 25 participants.

Table 2: Cross Tabulation of Respondents Based upon Enterprise Size and Management Decision (n=25)

	50	51-100	101-250	251-500	500+
Become actively involved with the Storage Policy decision	2	1	0	0	1
Discard data that had minimal value	2	0	0	1	2
Redefine the data that should be retained	0	0	0	1	5
Build a team that reviews the data being retained	4	0	1	0	5
Total	**8**	**1**	**1**	**2**	**13**

The data in the table shows that whether the enterprise is large or small the majority agree that a team should be assembled to review the data being retained. In this case, enterprise size versus management decision did not produce any new information. The extrapolated information from Table 2 shows that data with little or no value exists within the enterprise regardless of the size.

The analysis to this point has not involved senior management: chief executive officer (CEO) or chief information officer (CIO). Management often uses a top down or bottom up approach for storage management decisions. The intent of the following survey questions was to establish whether management was involved in the storage decisions. Understanding how the CEO and CIO participate in the decision making process provides an essential view of how the business mission supports the data storage policies and processes.

When the question was asked in the survey whether the CEO/CIO drive the business goals to all level of the organization, 73 percent of the participants said they did not. Additionally, 58 percent of the survey participants responded that even at the organizational level, the goals and mission of the organization are not defined to collect data with business value. Furthermore, 92 percent of the survey participants stated that the data being collected supports the business entities and goals of the organization.

To further expand the knowledge, another survey question asked whether stakeholders are involved in the storage decisions of the actual data collected. Out of the 26 participants, 50 percent stated that their stakeholders are involved. Additionally, 31 percent were not involved and 12 percent were asked to be involved but chose not to participate. The remaining 7 percent were not sure if there was stakeholder involvement. Understanding the relationship between the CEO/CIO and stakeholder participation would further expand my knowledge of how data is evaluated. Cross tabulating the two previous questions, survey questions 6 and 25, provided additional insight into their involvement within the enterprise. The data showed that approximately 37 percent of the participants stated that their enterprise lacked involvement from stakeholders, CEOs and CIOs as shown in Table 3.

Table 3: CEO/CIO Driven Business Goals versus Stakeholders Involvement in Data Storage Decisions

	Yes	No	I am not sure	They are part of the decision process but do not participate
Yes	57.14%	14.29%	28.57%	0%
No	47.37%	36.84%	0%	15.79%
Total	**50%**	**30.77%**	**7.69%**	**11.54%**

Within the enterprise some level of culture exists. Culture establishes the communication, decision and policy making processes throughout the enterprise. Culture can be well embedded and is not often easy to change or redefine. Figure 8 shows the survey participants response in regards to whether culture affects the data collection within the enterprise. If we consider the changes within the enterprise such as storage cost, capacity, enterprise mission, policies, and applications, one factor that often remains the same is the culture. Often what determines the value and defines the policies storing the data is cultural in nature. For instance, interviewee two

stated that often the wrong management team such as legal directs the information technology staff (personal communication, February 7, 2011). It is beneficial to understand the effects of culture upon data storage. The evidence is that approximately 77 percent of the survey participants state that the collected data is affected by enterprise culture.

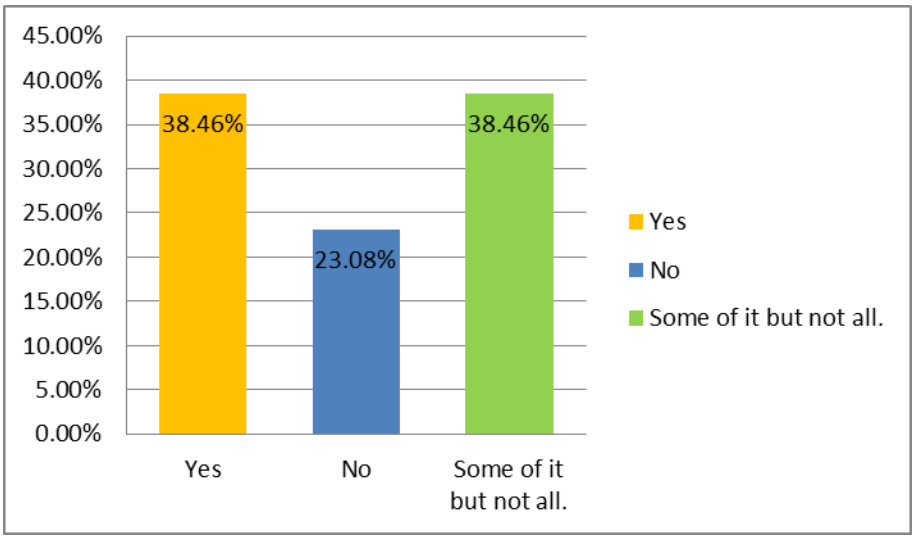

Figure 8: Data Collection due to Organizational Culture by Percentage of Respondents

Policies within the Enterprise

Policies dictate how the data value is defined, what is being stored, where it is stored at, and when it should be discarded. Policies are required to manage storage successfully. The survey asked the participants whether their enterprise have and use storage policies. Understanding policy implementation provides knowledge significance to this study. The survey question asked whether the enterprise have and use storage policies. The respondents were split, 50 percent of the participants stated that storage policies were implemented, while 31 percent stated they did not have storage policies implemented. Additionally, 19 percent stated storage policies were defined but not enforced. A number of factors affect best practice: culture, management and stakeholders. Ultimately, the goal of the study is to establish best practices that maximize storage within the enterprise and control storage expansion.

Data Valuing within Policies

In order to answer the questions of this study it became important to gain a higher level understanding of how data is valued within the enterprise. The survey asked the participants if data with no value was deleted within their enterprise. The survey showed that 72 percent of the participants do not delete data that has little or no value. Additionally, 8 percent responded that policies are defined to delete data with little or no value but are not enforced. It can be determined that 80 percent of the respondent's rules or policies do not manage data with little or no value. In these cases the defined policies did not inhibit storage growth.

Who is driving the storage policy decisions? Are the decisions being driven by the stakeholders, organization in general, business management, information technology management or other areas within the enterprise? The survey showed that 27 percent of the participants stated that the decision was made at the organizational level as shown in Figure 9.

I had a lengthy discussion with interviewee one in regards to storage, policies and data valuing. If the organizational level approach is taken, data is likely not being properly valued or maintained (personal communication, January 23, 2011). Furthermore, the data shows that 19 percent of the data management policy decisions are made by information technology management.

Organizational and stakeholder participation only occurs in 15 percent of the participant's enterprise. Additionally, 11 percent stated that the data management policies are being defined by management who do not understand the data.

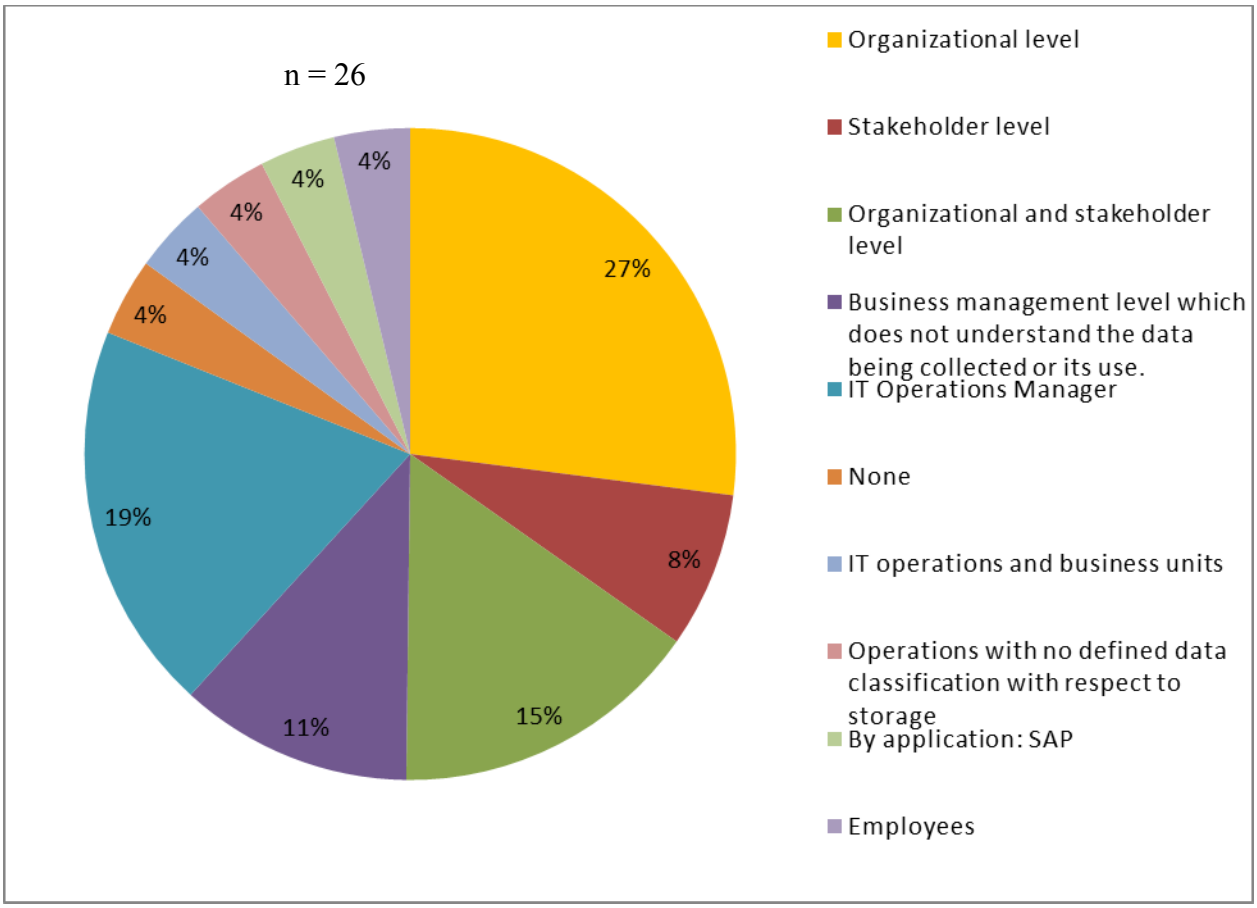

Figure 9: Enterprise Entities that Define Data Management Policies by Percentage of Respondents

Review of Policies

Within the enterprise it is important to understand if storage management policies are reviewed. Business mission and requirements change and it is important to understand whether storage management policies are dynamic or static. The survey data shows that 32 percent stated that their storage policies were reviewed and the stakeholders were involved while 36 percent stated the stakeholders were not. Additionally, 32 percent of the respondents stated that their storage management policies remain static.

Understanding what drives the creation of storage management within the enterprise is important from a policy and process level. Prior to defining storage policies a model should be defined that documents what data is going to be collected, valued and stored. It is important to

understand whether there is consistency in the storage decision making process. The survey participants were asked whether a data model was defined within their enterprise to ensure that a level of decision consistency existed. Figure 10 shows the results from the survey respondents.

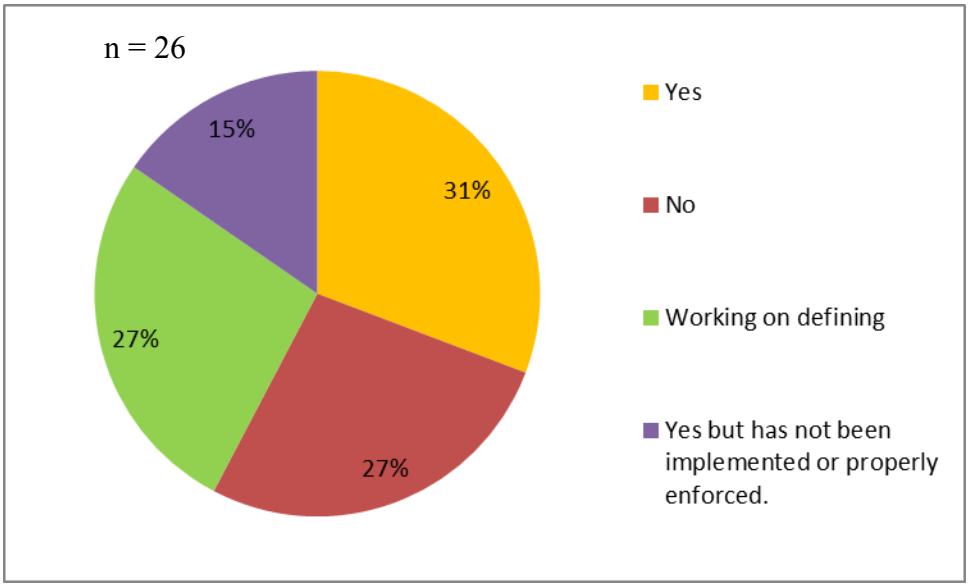

Figure 10: Implemented Data Model and Decision Consistency by Percentage of Respondents

From the data in Figure 10, 31 percent of the respondents stated that a model had been defined. The survey further showed that 69 percent of the participants have not or are working on defining a model.

The data model was discussed during the interviews. Interviewee one stated that metadata is the key. Furthermore it was stated that metadata should not be static but dynamic (personal communication, January 23, 2011). The data model ensures that context is defined and consistent across the enterprise. Interviewee two stated that metadata is misunderstood (personal communication, February 7, 2011). Standards often do not exist or metadata becomes too large and useless to be used effectively (personal communication, February 7, 2011).

Data Migration

Data migration is one method to reduce and control storage growth. Figure 11 shows the data from the survey. Migration to a lower storage tier or warehouse was selected by 42 percent of the participants. The survey question shows that 31 percent would delete the data with no or little value from primary storage. Interviewee two stated that from experience, 75 percent move everything during a data migration (personal communication, February 7, 2011). The other 25 percent try to understand the data at some level, such as the file types and data users (personal communication, February 7, 2011). The interview enforced the idea that the enterprise does not understand the data or its value.

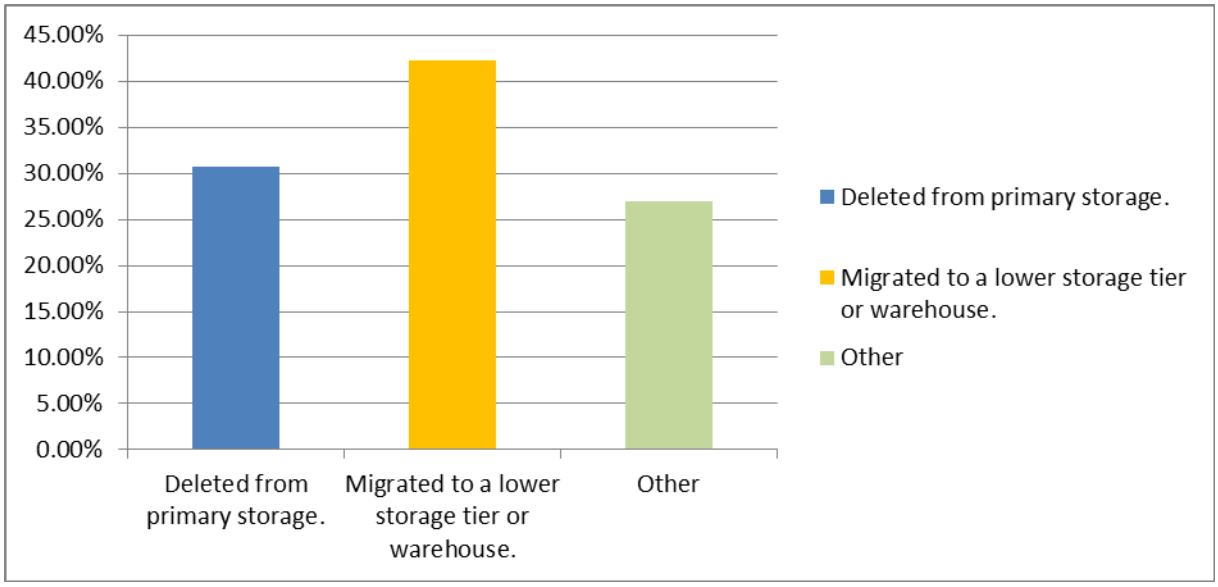

Figure 11: Data Movement to Regain Storage Space within the Enterprise

The last survey question asked the participants, how much primary storage could be regained if their storage management policies were based on data value. Figure 12 shows the results. The data shows that 39 percent of the participants stated that they could reclaim up to 50 percent of their primary storage if proper policies were implemented.

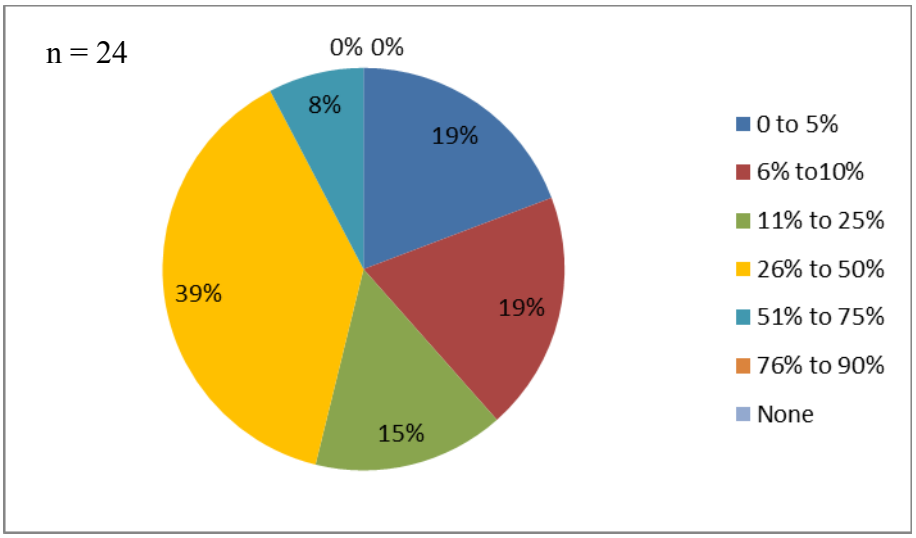

Figure 12: Regain of Primary Storage by Percentage of Respondents using Proper Policies

Performing a cross tabulation provides additional information into the amount of storage regained by organization size as shown in Table 4. Primary storage can be regained across all entities. The data showed that small and large enterprises will both benefit.

Table 4: Regain of Primary Storage by Enterprise Size and Primary Storage Reclaim (n=24)

	0-5%	6%-10%	11%-25%	26%-50%	51% - 75%
50 employees	1	1	0	4	1
51 – 100 employees	0	0	1	0	0
101 – 250 employees	0	0	1	0	0
251 – 500 employees	0	1	0	1	0
More than 500 employees	4	3	2	4	0
Total	**5**	**5**	**4**	**9**	**1**

Conclusion

The results of the data show that the enterprise does not managed data efficiently and effectively. The problem exists within the entire storage process from participation to implementation. The interviews have substantiated the survey finding in a number of areas such

as data valuing and policy involvement. The survey has shown that there is a lack of storage mission throughout the enterprise.

The survey revealed information that the enterprise can regain storage and control growth if policies are correctly defined and data is properly valued. This is lacking in the enterprise today maintaining data with little or no value. Furthermore, the survey data brings to light that the enterprise supports storage expansion rather than understanding the data that is being collected. It is clear that data can be assessed to better manage data storage including defining best practice policies.

Storage growth is not understood within the enterprise. Management approves storage expansion as a requirement to meet business mission but often the mission for storage is not defined or communicated. Additionally the survey has exposed that management does not understand the cost involved with storage; it is not even being presented to them. Chapter 5 will define best practices using the learned information from the interviews and survey.

Chapter 5 – Conclusions

The ability to collect and store data has changed how businesses make decisions and function every day. In many instances, data is power and financial security. What is often overlooked is the effects that storage growth has upon the enterprise. Defining best practice policies that control storage growth requires an understanding and participation on a number of levels. This provides the knowledge to answer the four questions of this study.

Understanding Storage within the Enterprise

The initial question was whether the enterprise understands their storage. The collected data indicates that the visibility of storage growth and cost is lacking at a number of levels. Management and information technology leaders understand that their data storage needs are expanding rapidly. Understanding storage refers to the data that is being captured and maintained including the cost to perform the storage tasks.

The survey has shown that management does not fully understand the associated cost to maintain data within the enterprise. The CEO/CIO typically is presented a yearly budget where the cost is simply unknown. Information technology, management and even culture can all be contributors to the cost notification method used within the enterprise. This is supported by the data collected from the survey. It is however evident that the enterprise understands that their requirement for storage is continually increasing year after year. It is true that the cost per gigabyte is decreasing; however the cost to manage the data is not understood. This is evident from the based upon the survey results.

The survey has substantiated that storage growth is understood, however the cost is not. Management is not presented with the correct information to make informed decisions in regards to storage expansion and data collection. If management understood the cost to store, manage

and move the data, then policies would be implemented to properly define storage levels. If I return to the survey, 40 percent of the participants stated that if the cost were understood, a team would be assembled to address the storage requirements. The enterprise has become its own worst enemy. Storage growth is being supported because the wrong information is being presented to the management levels.

Asking the survey participants what management would do if they understood their storage costs further substantiated that a failure to communicate exists within the enterprise. The survey clearly showed that management would become involved in the storage decision making process. It is an essential requirement that management is involved. This defines a best practice policy.

Can Best Practices be Defined?

Implementing best practice policies across an organization can be successfully implemented. The methodologies such as ILM and MDM exist and if implemented correctly can manage the data. From my research, using the interviews and surveys, the problem resides outside of the application solution. Best practice policies must begin with the enterprise mission. The mission has to be established prior to defining any data value.

In order to define the mission senior management support is required. Traditionally the mission is communicated to all levels of the organization. In defining the data storage mission, several factors need to be reviewed. Has senior management, information technology or stakeholders defined the organization's goal relating to data and storage? What has occurred in many enterprises is that information technology asks that you review and clean-up the data.

The survey showed that 57 percent of the participants stated that the mission and goals that align data with the data value does not exist. If I evaluate this with the 80 percent of the

participants that stated it is easier to expand storage than value the data, it becomes evident that proper storage policies are nonexistent or lacking within the enterprise. The enterprise mission is critical to the successful implementation of best practice policies. The mission of the enterprise often changes or are redefined; they are not static. If this is the case, then the mission change should drive policy review to ensure best practices are occurring.

One of the problems that the survey has shown is that only 50 percent of the participants stated that they had storage policies implemented. Additionally, 20 percent stated they existed but were not enforced. Policies have to be established and are essential in defining the storage requirements that driven from the enterprise mission. In the absence of storage policies there is no visibility into the data being within the enterprise. When storage policies are insufficient, then the data becomes independently maintained and silos of storage exist.

Best practices cannot be defined from the use of static storage policies. The mission changes within the enterprise so why shouldn't the storage policies? If we implemented an ILM solution and defined the best practice policies today, then they could be obsolete in a year or less. Policies need to be reviewed on a defined timeline. Again this is not an application issue but a management issue. Stakeholders have to be involved. The survey showed that 32 percent of the participants reviewed policies with stakeholder involvement while 68 percent did not. Of the 68 percent, 32 percent maintained static storage policies.

There is some value associated with all data. What is often not considered is that the value changes and can effect storage growth and best practice policies. Data that is dynamic today may be considered static tomorrow. Interviewee one shared a presentation which applied to medical data. The concept was that the data's value is every changing: static to dynamic or dynamic to static. This concept is often not understood within the enterprise. Because of this we

store data that could potentially have no or little value. The problem with best practice policies are multi-dimensional, however from my research it often begins with management and the mission.

Assessing the Data to verify it meets the Defined Mission of the Organization

Assessing or valuing the data is essential to defining best practices. We value many areas of the enterprise such as manufacturing processes, financials and the people. Why is data any different? In fact data is no different and should be assigned some value using methods that are acceptable to the enterprise. If data value is not understood, then best practices cannot be defined. Furthermore, if the data value changes it cannot be addressed and essentially is given the same worth.

From the survey it is apparent that a data model is often not defined for data consistency. Only 31 percent of the participants stated that a data model was defined. If the data model does not exist, then how are the decisions being made to define what data to capture and how it should be valued? It is apparent from the survey that most of the participant's enterprises have not defined a data model for decision consistency. Data models are essential to defining best practices.

Captured data has to support the mission of the business. Data with little or no value should be discarded to control storage growth. Consider that 80 percent of the participants stated that data with little or no value was not deleted. Furthermore, consider the effect of maintaining data with little or no value upon the storage year after year. Again, storage expansion is being promoted at rates that should not be occurring.

The reason data with no value is maintained is simple; it is not understood. It is a challenging and daunting task to define the data value within the enterprise. Data is what makes

the enterprise unique but the risk in maintaining data should be considered and addressed within the storage policies. Data that is not valued properly and has little or no value continues to be maintained due to the storage policies inability to address it. The survey shows that data with no value is maintained. This is substantiated by the fact that 38 percent of the participants could regain between 25 percent and 50 percent of their storage if data with little or no value was deleted.

What would you do if storage could no longer be expanded? The majority of the survey participants stated that they would value their data if in this situation. Yet the enterprise continues to expand their storage rather than value the data. The enterprise stores data because it can. Tools have become very efficient at storing data for analytics. The reason is simple; the collected and stored data may provide value or be an asset at some time. Consider that 96 percent of the survey participants stated that this is why the data they collect is maintained.

Collecting and storing data does not mean that it supports the enterprise mission or promotes revenue. The survey showed that 46 percent of the participants stated that the collected data supported business revenue while 34 percent said it did not. Additionally, 20 percent of the participants stated that they were not sure who used the data for what. The enterprise does not understand data value and what data brings value to it. It is apparent from the survey that there is approximately a 50 percent divide of whether the captured and maintained data supports the enterprise positively. Data value changes and policies need to be redefined to accommodate change. Storage expansion can be controlled by policies but they are dependent upon the data's value being defined and redefined as the enterprise changes.

Defining Best Practice Policies

Defining best practices that will reduce storage growth requires a framework to be established. The lifecycle management methodologies do not offer this at a simplistic level and do not focus on controlling storage growth but often relies on moving data to the correct tier. In order to be successful, a framework that addresses data management and valuing is required. The objective of the framework is solely on understanding the data prior to any policies being defined. Successfully working through the framework will provide best practice policies that can be applied to the enterprise.

Storage Quadrant Framework (SQF) is based upon four quadrants that are focused on educating the enterprise where their storage knowledge is lacking. No two enterprises are alike in regards to storage, so the policies should not be either. The framework has to be flexible enough to accommodate this. Figure 13 shows SQF. The concept of SQF is based upon the enterprise's management and stakeholder involvement.

The strength of SQF lies in the relationships between the quadrants ultimately defining best practice policies for the enterprise. Each quadrant covers a specific task using a likert scale approach as a guideline. The goal of the quadrants is not to enforce a policy but to build best practices based upon storage requirements. The directional arrows on the outside of the four quadrants show that any quadrant can be revisited as change is evident within the enterprise. Defining best practices cannot be successfully performed if the collected data is considered static. When all quadrants work together the enterprise mission is satisfied and revenue generation is supported. Ultimately, data is collected to drive business profitability and compliancy.

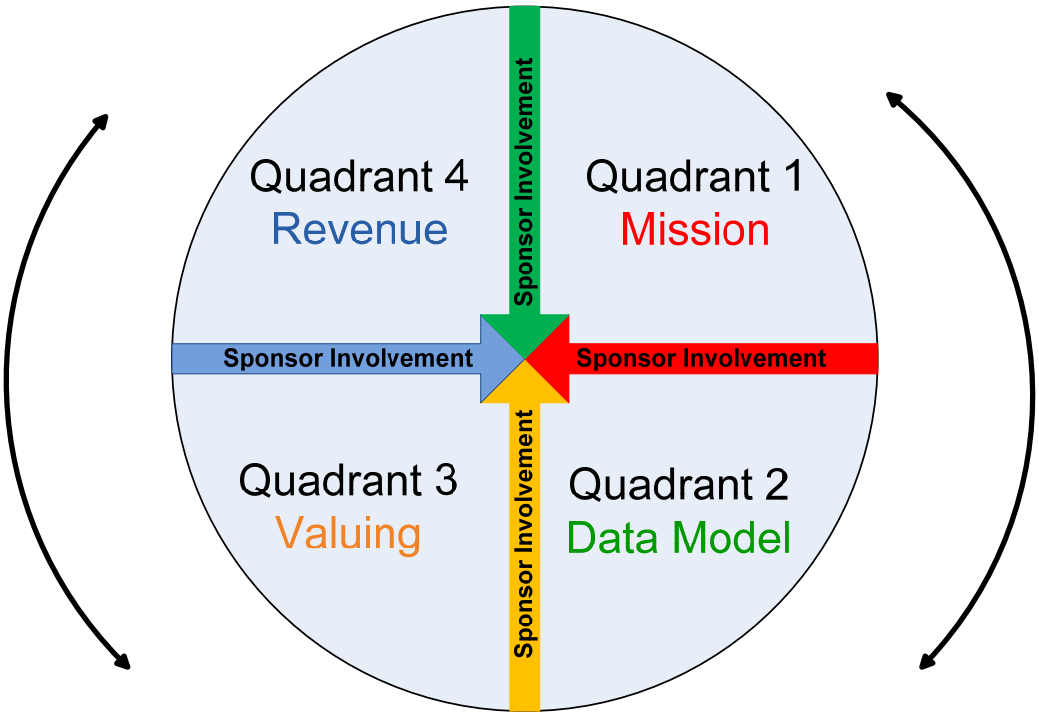

Figure 13: Storage Quadrant Framework (SQF)

Likert Scale for SQF

The likert scale defines where the enterprise lacks involvement or understanding of the storage or data. Using a simple likert scale with questions based upon the quadrant allows two objectives to be met. The first objective is the visibility into where the enterprise lacks understanding. The likert scale in table 5 and Figure 14 provides visibility into where the enterprise is deficient. The second objective is to define the best practice policy based upon the data for the enterprise that supports reducing storage growth. It is imperative to understand that storage growth refers to all data, not only new.

The likert scale used for each quadrant is depicted in Table 5. A color has been associated with the scaling to bring visibility to the requirements that need to be addressed. I have found in my experiences that a likert scale alone does not often show urgency and association to a color is a better visual indicator. Figure 14 depicts the scale with color.

Table 5: Likert Ratings and Definitions for storage quadrant framework

Scale	Color	Definition
1 - 3	Red	Does not exist
4 - 5	Orange	In Discussion and needs to be defined
6 - 7	Yellow	Exist but non-functional
8 - 9	Lime Green	Functioning but requires change
10	Green	Meets enterprise storage requirements

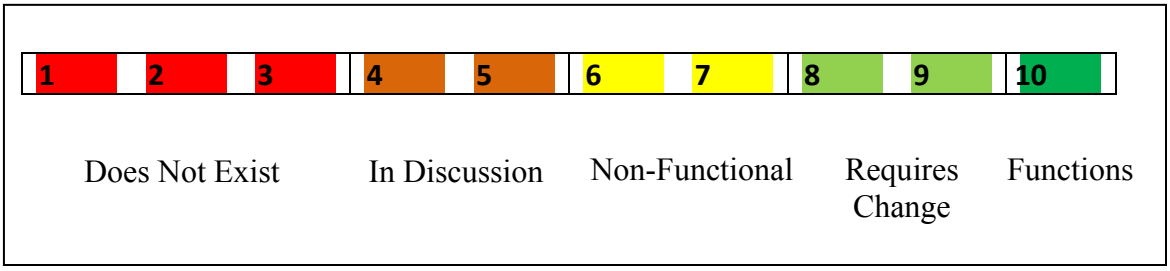

Figure 14: Storage Quadrant Framework Likert Scale with Color Definition

Defining the Quadrants to Reduce or Control Storage Growth

Quadrant 1: Mission

A defined mission is essential to any enterprise. The mission defines the goals and objectives of the enterprise and often establishes the culture within. The storage mission establishes that all management levels, stakeholders and employees are working based on the same policies and rules. The survey has shown that management lacks understanding of the real cost of storage that it costs the enterprise. It was further established that if management understood the cost of storage, then a team would be assembled to review the data being captured. The mission establishes that this criterion is met prior to proceeding to quadrant two. The following questions are answered in quadrant one.

- Is management involved with the storage decisions?
- Are the stakeholders and middle management involved with the storage decisions?
- Does management understand the “true” cost of storage at all levels?
- Is the enterprise data collection mission established?
- Is the enterprise data storage mission established?
- Is it established that the collected data fulfills the business mission of the enterprise?
- Is the mission communicated and supported through all levels of the organization?

The questions above establish that all entities within the enterprise are involved with the storage decision process. If we do not establish a management structure, the captured data is duplicated and likely siloes are formed. When all of management is involved, then there is visibility across all entities which promote data sharing thus reducing duplication. Furthermore, costs are established and controlled at all levels of the enterprise. Collecting and analyzing data is not the problem. The collection of data with no viable use or an expired shelf life is. David Loshin stated that without senior management support it would be difficult to execute any enterprise activity (Loshin, 2009).

This is why it is essential to establish and define a data collection and storage mission within the enterprise. This is a best practice. The definitions are supported by all levels of management and supported by rules, policies and procedures. In reviewing the survey, it is clear that storage solutions have been implemented which do not solve the storage growth challenge. Clearly policies are defined but the data continues to grow at significant rates. The single most important lacking factor is management involvement to understand the affects that the collected data places upon the enterprise. The survey data has shown that 73 percent of the participants stated that the

business objectives are not driven to all levels of the enterprise, which is promoting data storage that provides no value. Good data collection and best practice policies begin with the mission.

Quadrant 2: Data Model

Quadrant two establishes the data model. The data model ensures that the entire enterprise is working from the same set of storage rules. Clearly different entities have different requirements, but it is imperative that all work off of the same data model. The data model establishes that storage decisions are based upon the same criteria. This is a critical quadrant for the SQF. In order to begin quadrant two the mission has been defined and is supported.

The primary objective of the data model is to classify the data. Data classification is required prior to defining value. There is multiple levels of data classification that needs to be considered which affects the enterprise. The following questions are answered in quadrant two.

- Is a data model defined across the enterprise?
- Does the data model support the mission of the enterprise?
- Is the right data being collected?
- Is it understood where the data is maintained including its storage requirements?
- Are the requirements for the stored data defined and understood?
- Is it understood whether the data is dynamic or static?
- Is the data classified by use, compliancy and business need?

A data model has to be defined that supports the storage mission of quadrant one. It is impossible to value data in quadrant three unless it can be understood and classified.

The data model establishes that the data has a purpose and provides the enterprise some value. The value is not defined but quadrant two establishes that the right data is being collected and maintained. Quadrant two can bring visibility to management showing that the correct data

may not be collected indicating a deficiency. Quadrant two can also provide insight to the data being collected that does not meet the enterprise storage or business mission.

Essential questions in quadrant two address the ability to determine whether the data changes and if it is required for compliancy. The enterprise must understand whether the data they are collecting and storing is static or dynamic. If the data is dynamic, then questions have to be further answered that defines why and how it is addressed. A definition has to be created within the model allowing static and dynamic data to be differentiated. Storage growth cannot be contained or reduced if the data is not understood at this level.

It becomes clear that storage policies cannot be properly defined if this level of understanding does not exist. Treating all data as static or dynamic is not realistic. If a lack of knowledge exists, then data is misunderstood and all data is maintained. Data may be moved to a new tier and the primary issue remains and storage continues to expand. The data model has to clearly define and understand if the data does or does not change.

Quadrant two has to define and accommodate data compliancy. This clearly affects quadrant three and the data has to be understood. Data that meets compliancy becomes a risk when it reaches a defined period of time and the proper entities with the enterprise have to understand that risk. Information technology and/or management alone should not make the storage decisions.

The survey data showed that 69 percent of the participant maintained all or most of their data because of compliancy or governmental requirements. It is hard to believe that all data is required to meet these types of requirements. What I believe, drives this is a lack of education and definition of the data. The data model not only educates management and information technology because all entities are involved. It further establishes the definition to ensure the

right data is collected and maintained. The data model is a best practice policy to reduce and control storage growth.

Quadrant 3: Valuing

Quadrant two established the data model ensuring that the correct data to fulfill the mission of the enterprise is being captured and stored. Quadrant three defines the value of that data. Data valuing is essential in defining best practice policies. The value determines how the data is stored and when it has reached its life expectancy. Even data that meets compliancy has a defined value that establishes when it can be discarded. Defining data value is not easy and the questions for this quadrant have to rely upon knowledge, mission and management support. It is imperative that the data is understood and valued supporting the enterprise mission. The following questions are asked in quadrant three.

- Are the correct entities and management involved to value the data?
- Is a method established such as metadata to define the value across all data?
- Does the method consider all data types supporting value definition and policy enforcement?
- Can the method be updated dynamically?
- Is a definition created that establishes data value such as critical, high, average, and minimal?
- Is a definition defined that establishes time periods of storage for the data value?
- Is a value definition defined establishing how the data value promotes the enterprise mission?

Valuing the data requires definitions to be established for the enterprise. Different value definitions cannot exist within the enterprise and support the same mission or storage goals. The

survey showed that 46 percent of the participants stated that the collected data provided revenue to the enterprise. At the same time the survey again showed that data with little or no value could regain up to half of their primary storage for over 50 percent of the participants.

If the enterprise collects and maintains all data, it will undoubtedly meet the business requirements while at the same time cause storage to expand uncontrollably. In order to implement a value definition, a method has to be established. The valuing definitions have to provide a method that ensures data with no value can be easily determined. The method has to be defined and dynamic to allow the data to be properly valued. What is often not considered is how the value of the data changes. Establishing that the method can be dynamically changed provides an effective tool when applying value definition. All data is not valued the same nor should it have the same life requirements.

The issue is not the lifecycles that are implemented within the enterprise do not address storage, but that the data value is not understood and application policies are not properly defined. Data value is not evaluated at all levels of the enterprise. All stakeholders including senior management must be involved and a live participant. The survey data shows that only 26 percent participated in data valuing at the organizational level.

The survey further revealed that 20 percent of the data valuing decision was made by information technology management. In my experience, information technology will maintain more of the data than less to protect the enterprise and ensure they can provide answers to the entities when required. This is not the correct storage approach and again promotes expansion. Other enterprise leaders that are involved with defining the storage policies, such as business management or stakeholders, are likely not qualified to define data value.

Definitions have to be established that define the level of value and the period of time associated with value. The method has to support the valuing system in order to regain and control storage growth. It is essential that the valuing approach is documented and supports the enterprise mission and data model. Valuing data is a best practice to control storage growth and regain storage space.

Quadrant 4: Revenue

Data that is captured and stored should drive enterprise revenue in some form. When quadrant four is reached it should be evident that the data being collected and valued provides or supports revenue. The survey data showed that 58 percent of the participants stated that the data being collected does not have business value. Business value translates into revenue whether it is product design or a service being offered. If quadrant four is reached and the questions below cannot be answered in the green area of the likert scale, a significant problem exists.

- Evidence exists that the collected data supports the mission and enterprise revenue?
- The defined data and data value protects the enterprise revenue?
- The collected data supports enterprise success?
- Management understands the effects of the collected data upon the revenue?
- Data that does not support revenue is discarded?

If the collected data supports the enterprise mission and is valued appropriately quadrant four should be fulfilled. Quadrant four is a check and balance that ensures the previous three quadrants were completed successfully. It is difficult to reach this stage and determine that re-evaluation is required because the above questions cannot be successfully answered. If the correct data is collected and stored, the revenue goals of the organization are supported.

Revenue and data are not always aligned. Data is collected for many reasons but if it does not drive revenue why is it maintained? Data that is maintained for compliancy such as emails may be an area that could contest this but the fact is that email is a business tool that produces transactions. All data does not support revenue, non-revenue data needs to be evaluated and understood. Simply put, the data has to support the mission of the enterprise or it has no value and should not be stored or collected. Storage growth is driven by data that does not support revenue. Best practice is defining policies that support enterprise revenue.

Applying the Quadrants to the Likert Scale

We have become too accustom at performing analysis using scales and automated analytics. Determining the mission and value of the data requires a higher understanding than just allowing a policy to define it. In the case of SQF, the purpose of the likert scale is to provide a visual gauge of where the enterprise lies within the quadrants. If the enterprise resides in the red, orange or yellow area of the likert scale, issues have to be addressed and advancing to the next quadrant should not occur. Clearly different entities within the enterprise will score differently; this is expected. The likert scale is a tool to initiate and promote conversation in regards to the data being collected and stored. Data value cannot be defined from a scale; it requires communication. Best practice policies require communication at all levels of the enterprise. A lack of communication will ensure the wrong data, duplicate data or data with little or no value is being stored. The enterprise is driving non-essential storage growth because the data is not understood at all levels. SQF brings visibility and promotes open communication across all levels to solve of the enterprise to solve this problem.

Conclusion

The study has shown that storage can be recovered and data is being maintained and collected with little to no value. The study has further shown that data understanding and enterprise mission are not communicated throughout the enterprise.

We have become data junkies and collect data because we can. The cost per gigabyte continually decreases while the amount of gigabytes required continues to increase. Unfortunately the survey has shown that it is easier to expand storage rather than value data. Additionally, management does not understand the cost of storage across the enterprise. The enterprise is its own worst enemy in driving the ever increasing storage requirements. It takes time and resources to evaluate data, but storage growth and cost will outpace the cost per gigabyte savings and management will realize what the true cost to manage, migrate and maintain the data is.

In addition to the lack of management participation in the storage decisions, lifecycles have been implemented to control the data and ensure it meets the business requirements. The results of this study substantiate that data with little or no value is maintained and the solution is to either expanded or moves it to a different storage tier. This is by no means failure of the lifecycle but a reflection on the implementation. Policies within the lifecycle dictate how the data is handled. Knowledge of the “true” data value is however unknown so the enterprise continues to store instead of evaluate and properly define the best practice policy. The enterprise defined policies support data storage, not data deletion due to a lack of mission, data model and valuing.

Best practice policies require a practical framework outside of the lifecycle. The SQF framework was designed based upon the responses from the survey, interviews and literature review to provide the enterprise visibility into why best practices are not defined and storage

continues to expand. The enterprise is missing the most primitive idea that data storage is a project and should be treated as one. Management support is required throughout the entire process. The strength of SQF is participation and the sharing of knowledge by all entities and stakeholders throughout the process. It is very evident from Figure 13 that the circle represents strength which increases at the center of the framework driven by stakeholder participation.

There is however an opposing view that would suggest why data with little or no value should be maintained. In today's environment, e-discovery is a tool that provides the party's within a litigation the means to acquire information that can support their case. For example, emails and chats that seem to have little or no value may prove innocence or negligence, affecting the legal outcome and ultimately affecting the enterprise. This topic area adds the dimension of risk to the deletion of data that is perceived to have no value. Furthermore, the data being considered to delete or maintain is not due to a compliancy or legal hold requirement.

In these cases, management may decide to move and store data with little or no value to a lower tier such as a data warehouse. The senior management team has to further understand whether the risk level of data deletion is acceptable, rather than continuing to store it. Understanding and defining the acceptable level of risk within the enterprise weighs heavily in this decision making process. It becomes clear that data with no value could potentially affect the outcome due to the unknown risk. The topic of risk, data deletion and e-discovery provides a further study area that could expand this thesis. Even in the case of e-discovery and data deletion, best practices play an essential role in the decision making process.

Best practice policies are lacking within the enterprise for the simple fact that the data being collected is not understood coupled with the fact that the financial burden associated with

it is unknown. The SQF brings visibility to these areas within the enterprise. Best practices come from understanding the enterprise mission, data being collected and defining the value.

References

Agrawal, N., Bolosky, W., Douceur, J., Lorch, J. (2007). A five-year study of file-system metadata. *Trans. Storage* 3, 3, Article 9 (October 2007). DOI=10.1145/1288783.1288788 http://doi.acm.org.dml.regis.edu/10.1145/1288783.1288788

Carter, P., & Green, G. (2009). Networks of contextualized data: a framework for cyberinfrastructure data management. *Communications of the ACM*, *52*(2), 105-109. Retrieved from http://web.ebscohost.com.dml.regis.edu/ehost/pdfviewer/pdfviewer?vid=7&hid=12&sid=74048fbf-de06-4ad8-90d7-284dbc886a2d%40sessionmgr11

Chenyang Lu, Guillermo A. Alvarez, and John Wilkes. (2002). Aqueduct: online data migration with performance guarantees. In *Proceedings of the 1st USENIX Conference on File and Storage Technologies* (FAST '02). USENIX Association, Berkeley, CA, USA, Article 21. Retrieved from http://portal.acm.org.dml.regis.edu/citation.cfm?id=1083323.1083351&coll=DL&dl=GUIDE&CFID=4530812&CFTOKEN=78016080

Ciroth, A. (2009). Cost data quality considerations for eco-efficiency measures. *Ecological Economics*, *68*(6), 1583-1590. doi:10.1016/j.ecolecon.2008.08.005. Retrieved from http://www.sciencedirect.com.dml.regis.edu/science?_ob=ArticleListURL&_method=list&_ArticleListID=1520955310&_sort=r&_st=13&view=c&_acct=C000055361&_version=1&_urlVersion=0&_userid=1922016&md5=4ec7ecaa2be527b4d1531ceb116a4565&searchtype=a

Detlor, B. (2009). Information Management. *International Journal of Information Management, 30*(2010), 103-108. Retrieved from

http://www.sciencedirect.com.dml.regis.edu/science?_ob=MImg&_imagekey=B6VB4-4NP3P7S-1-1&_cdi=5916&_user=1922016&_pii=S0268401206001435&_origin=search&_coverDate=08%2F31%2F2007&_sk=999729995&view=c&wchp=dGLzVzb-zSkzS&md5=3891eb1f4bc0a25a2676965fc28b833a&ie=/sdarticle.pdf

Eckerson, W. (2002). Data quality and the bottom line. Retrieved from http://download.101com.com/pub/tdwi/Files/DQReport.pdf

Horwit, E. (2008). Data lifecycle management no panacea. SearchCIO.com Retrieved from http://searchcio.techtarget.com/news/1323186/Data-lifecycle-management-no-panacea

Information Management. (2010). Data Multiplies as Storage Costs Fall. *Information Management (15352897)*, *44*(3), 4. Retrieved from Business Source Complete database. Retrieved from http://web.ebscohost.com.dml.regis.edu/ehost/pdfviewer/pdfviewer?vid=5&hid=119&sid=d1f9cd62-d1b2-4dd2-a6af-a29ed86eb0f3%40sessionmgr111

Information Management Journal. (2009). Data growth driving archival efforts." *Information Management Journal* 41.2 (2007): 14. *Academic Search Premier*. EBSCO. Web. 3 Oct. 2010. Retrieved from http://web.ebscohost.com.dml.regis.edu/ehost/detail?vid=3&hid=11&sid=31fbc407-86f0-4bc8-b86e-f8806494421c%40sessionmgr4&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=aph&AN=24831401

Inmon,W., O'Neil,B., Fryman, L. (2008). *Business metadata.* Burlington, MA: Kaufmann Publishers.

Kristin Weber, Boris Otto, and Hubert Ősterle. (2009). One size does not fit all- a contingency approach to data governance. *Journal of Data and Information Quality* 1, 1, Article 4 (June 2009), 27 pages. DOI=10.1145/1515693.1515696. Retrieved from http://doi.acm.org.dml.regis.edu/10.1145/1515693.1515696

Loshin, D. (2009). *Master data management.* Burlington, MA: Morgan Kaufmann

Lundell, B., McKnight, J. (2009). Enterprise storage priorities emphasize information and infrastructure efficiency. Retrieved from http://www.enterprisestrategygroup.com/2009/01/esg-research-brief-enterprise-storage-priorities-emphasize-information-and-infrastructure-efficiency/Maes, R. (2007). Data and reality: a plea for management realism and data modesty. *Ubiquity* Apr. 2007, 1 pages. DOI=10.1145/1241854.1241855 http://doi.acm.org.dml.regis.edu/10.1145/1241854.1241855

Mearian, L. (2010). Data growth remains IT's biggest challenge, Gartner says. *Computerworld.* Nov, 2010. Retrieved from http://www.computerworld.com/s/article/9194283/Data_growth_remains_IT_s_biggest_challenge_Gartner_says

Mosley, M., Brackett, M., Earley, S., Henderson, D.(2009). *The dama guide to data management body of knowledge.* Bradley Beach, NJ: Technics Publications, LLC.

Petrocelli, T. (2006). *Data protection and information lifecycle management*. Stoughton, MA: Pearson Education Inc.

Rajasekar, A., Moore, R., Wan, M., Schroeder, W. (2010). Policy-based distributed data management systems. *Journal of Digital Information, 11(1).* Retrieved from http://journals.tdl.org/jodi/article/view/756/644

Russom, P. (2006). Taking data quality to the enterprise through data governance (report excerpt). Retrieved from http://tdwi.org/articles/2006/05/09/taking-data-quality-to-the-enterprise-through-data-governance-report-excerpt.aspx

Smalltree, H. (2006). Data governance requires checks and balances, gartner says. Retrieved from http://searchdatamanagement.techtarget.com/news/1230521/Data-governance-requires-checks-and-balances-Gartner-says

Socci, D. (2004). The make or break role of information lifecycle management. *Virtualization Journal*. Retrieved from http://virtualization.sys-con.com/node/47517

Swartz, N. (March-April 2007). Data growth driving archival efforts. (UP FRONT: News, Trends & Analysis). *Information Management Journal*, 41, 2. p.14(1). Retrieved April 05, 2009, from Academic OneFile via Gale: http://find.galegroup.com.dml.regis.edu/itx/start.do?prodId=AONE

Tallon, P. (2010). Understanding the dynamics of information management costs. *Communication of the ACM,* 53, 5. p.122-125. Retrieved from http://web.ebscohost.com.dml.regis.edu/ehost/pdfviewer/pdfviewer?vid=7&hid=105&sid=7e455b1c-7edf-42a8-8745-e71dc6b24c2a%40sessionmgr112

Teachey, D. (2009). MDM: it's not just about the technology. Retrieved from http://mdmbook.com/?p=68

Vengerov, D. (2008). A reinforcement learning framework for online data migration in hierarchical storage systems. *Journal of Supercomputing*, 43(1), 1-19. doi:10.1007/s11227-

007-0135-3. Retrieved from http://web.ebscohost.com.dml.regis.edu/ehost/pdfviewer/pdfviewer?hid=17&sid=574555fd-e509-413d-8082-bfb7543efe03%40sessionmgr12&vid=4

Ventana Research. (2009). Closing the data divide between business and it. Retrieved from http://www.ventanaresearch.com/whitepapers/_pdfs/Ventana_Research_Closing_the_Data_Divide_Between_Business_and_IT_White_Paper_2009.pdf

Vijay Khatri and Carol V. Brown. 2010. Designing data governance. *Commun. ACM* 53, 1 (January 2010), 148-152. DOI=10.1145/1629175.1629210. Retrieved from http://doi.acm.org.dml.regis.edu/10.1145/1629175.1629210

Appendix A

1. Profession
    a. Storage Expert
    b. IT Administrator
    c. IT Management
    d. Business Management
    e. CIO
    f. CEO
    g. Systems Engineer
    h. Other

2. Organization Size
    a. 50 employees
    b. 51 – 100 employees
    c. 101 – 250 employees
    d. 251- 500 employees
    e. More than 500

3. How often is your organization expanding its storage?
    a. Every 1 - 3 months
    b. Every 3 - 6 months
    c. Every 6 - 9 months
    d. Yearly
    e. Never

4. Does your organization classify its data to delineate which data has the most value?
    a. Yes
    b. No

5. Are the stakeholders involved in the data storage decision of the actual data collected?
    a. Yes
    b. No
    c. I am not sure
    d. They are part of the decision process but choose not to participate.

6. Does your organization have and use storage policies?
    a. Yes
    b. No
    c. They exist but are not enforced.

7. Are the storage policies reviewed periodically and are the stakeholders involved to ensure that the collected data is valuable to the organization?
    a. Yes, they are reviewed and the stakeholders are involved.
    b. No, they are not reviewed and remain fairly static.
    c. They are reviewed periodically but the stakeholders are not involved.

8. How does your organization handle the need for expanding storage requirements?
    a. Purchase more storage.
    b. Move files to cheaper storage or off-line.
    c. Delete old or unneeded files to recover storage space.
    d. Other

9. Is it easier for your organization to add storage rather than define or re-define data value within the storage policies?

    a. Yes it is easier because we do not have the resources to define data value.

    b. Yes but I am not sure of the reasons why.

    c. No we review the data value and manage the existing data based upon its value prior to adding storage.

    d. Yes because our storage policies do not change.

10. Has your organization implemented a data lifecycle management (DLM) or information lifecycle management (ILM) solution?

    a. Both

    b. Data Lifecycle Management solution

    c. Information Lifecycle Management solution

    d. Other

11. If your organization uses a data lifecycle management or information lifecycle management solution do they address the data properly or is there a conflict between solution implementations?

    a. They are different solutions and are implemented properly to address different problems.

    b. Both solutions are managed by different groups and there is no visibility whether a conflict exists.

    c. There is a conflict.

    d. Other

12. Does your organization clearly define the goals and mission of the organization to ensure that data with business value is collected?

    a. Yes

    b. No

13. If data is determined to have no or little value is there a “shelf life” that determines it can be:

    a. Deleted from primary storage.

    b. Migrated to a lower storage tier or warehouse.

    c. Other

14. Do your data management policies consider the “data value” when originally storing the data?

    a. Yes

    b. No

    c. Unsure

15. If you could no longer expand your storage and had to delete some to accommodate growth, would you make it a point to understand the value of the data to ensure the right data is maintained?

    a. Yes

    b. No

    c. I would take the FIFO (first in - first out) approach.

16. The policies that are defined to manage data look at value from the:

    a. Organizational level

    b. Stakeholder level

    c. Organizational and stakeholder level

    d. Business management level which does not understand the data being collected or its use.

    e. IT Operations Manager

    f. None

    g. Other

17. If you properly defined your storage policies based on data value and discarded data with no or little value you could regain_____ of primary storage:

    a. 0 to 5%

    b. 6% to10%

    c. 11% to 25%

    d. 26% to 50%

    e. 51% to 75%

    f. 76% to 90%

    g. None

18. The storage at your organization increases by:

    a. 0 to 10% annually

    b. 11% to 25% annually

    c. 26% to 50% annually

    d. 51% to 100% annually (doubles)

    e. More than 100% annually

    f. It is decreasing annually

19. Is the value of the data measureable either in revenue, product or future offerings?

    a. Yes

    b. No

    c. Not sure who uses the data for what.

20. Does your organization consider the collected data to be an asset and maintains all or most because it believes it has or will have some value?

    a. Yes

    b. No

    c. Unsure

21. Does your organization considers the collected data to be an asset and maintains all or most because of compliance or governance requirements?

    a. Yes

    b. No

    c. Unsure

22. Is the data that is collected due to an organizational culture?

    a. Yes

    b. No

    c. Some of it but not all.

23. Has a data model been defined to ensure decision consistency across the organization?

    a. Yes

    b. No

    c. Working on defining

    d. Yes but has not been implemented or properly enforced.

24. Does the CEO/CIO drive the business goals to all level of the organization promoting data storage that provides the business value?

    a. Yes

    b. No

25. Are data rules defined to delete data with no value?

    a. Yes

    b. No

    c. Yes they are defined to delete data but not enforced.

26. Does the organization consider the cost of storing, analyzing and moving data?

    a. The cost is hidden in the information technology budget.

    b. The cost has never been presented.

    c. No because it is a business requirement and cost does not matter.

    d. No one really understands the cost of storing data.

27. If the proper levels of management understood all the cost of storing data would they (selected the #1 reason that applies):

    a. Become actively involved with the Storage Policy decisions.

    b. Discard data that had minimal value.

    c. Redefine the data that should be retained.

    d. Build a team that reviews the data being retained and evaluate the value at some defined period of time.

28. Does the stored data support the key business entities and goals of the organization?

    a. Yes

    b. No